Graph-Structured Representations for Visual Question Answering
Damien Teney, Lingqiao Liu, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1-9
This paper proposes to improve visual question answering (VQA) with structured representations of both scene contents and questions. A key challenge in VQA is to require joint reasoning over the visual and text domains. The predominant CNN/LSTM-based approach to VQA is limited by monolithic vector representations that largely ignore structure in the scene and in the question. CNN feature vectors cannot effectively capture situations as simple as multiple object instances, and LSTMs process questions as series of words, which do not reflect the true complexity of language structure. We instead propose to build graphs over the scene objects and over the question words, and we describe a deep neural network that exploits the structure in these representations. We show that this approach achieves significant improvements over the state-of-the-art, increasing accuracy from 71.2% to 74.4% in accuracy on the "abstract scenes" multiple-choice benchmark, and from 34.7% to 39.1% in accuracy over pairs of "balanced" scenes, i.e. images with fine-grained differences and opposite yes/no answers to a same question.
**********

Physics Inspired Optimization on Semantic Transfer Features: An Alternative Method for Room Layout Estimation
Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, Li Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 10-18
In this paper, we propose an alternative method to estimate room layouts of cluttered indoor scenes. This method enjoys the benefits of two novel techniques. The first one is semantic transfer (ST), which is: (1) a formulation to integrate the relationship between scene clutter and room layout into convolutional neural networks; (2) an architecture that can be end-to-end trained; (3) a practical strategy to initialize weights for very deep networks under unbalanced training data distribution. ST allows us to extract highly robust features under various circumstances, and in order to address the computation redundance hidden in these features we develop a principled and efficient inference scheme named physics inspired optimization (PIO). PIO's basic idea is to formulate some phenomena observed in ST features into mechanics concepts. Evaluations on public datasets LSUN and Hedau show that the proposed method is more accurate than state-of-the-art methods.
**********

Local Binary Convolutional Neural Networks
Felix Juefei-Xu, Vishnu Naresh Boddeti, Marios Savvides; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 19-28
We propose local binary convolution (LBC), an efficient alternative to convolutional layers in standard convolutional neural networks (CNN). The design principles of LBC are motivated by local binary patterns (LBP). The LBC layer comprises of a set of fixed sparse pre-defined binary convolutional filters that are not updated during the training process, a non-linear activation function and a set of learnable linear weights. The linear weights combine the activated filter responses to approximate the corresponding activated filter responses of a standard convolutional layer. The LBC layer affords significant parameter savings, 9x to 169x in the number of learnable parameters compared to a standard convolutional layer. Furthermore, the sparse and binary nature of the weights also results in up to 9x to 169x savings in model size compared to a standard convolutional layer. We demonstrate both theoretically and experimentally that our local binary convolution layer is a good approximation of a standard convolutional layer. Empirically, CNNs with LBC layers, called local binary convolutional neural networks (LBCNN), achieves performance parity with regular CNNs on a range of visual data sets (MNIST, SVHN, CIFAR-10, and ImageNet) while enjoying significant computational savings.
**********

Designing Effective Inter-Pixel Information Flow for Natural Image Matting
Yagiz Aksoy, Tunc Ozan Aydin, Marc Pollefeys; Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 29-37
We present a novel, purely affinity-based natural image matting algorithm. Our method relies on carefully defined pixel-to-pixel connections that enable effective use of information available in the image and the trimap. We control the information flow from the known-opacity regions into the unknown region, as well as within the unknown region itself, by utilizing multiple definitions of pixel affinities. This way we achieve significant improvements on matte quality near challenging regions of the foreground object. Among other forms of information flow, we introduce color-mixture flow, which builds upon local linear embedding and effectively encapsulates the relation between different pixel opacities. Our resulting novel linear system formulation can be solved in closed-form and is robust against several fundamental challenges in natural matting such as holes and remote intricate structures. While our method is primarily designed as a standalone natural matting tool, we show that it can also be used for regularizing mattes obtained by various sampling-based methods. Our evaluation using the public alpha matting benchmark suggests a significant performance improvement over the state-of-the-art.
**********

Face Normals "In-The-Wild" Using Fully Convolutional Networks
George Trigeorgis, Patrick Snape, Iasonas Kokkinos, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 38-47
In this work we pursue a data-driven approach to the problem of estimating surface normals from a single intensity image, focusing in particular on human faces. We introduce new methods to exploit the currently available facial databases for dataset construction and tailor a deep convolutional neural network to the task of estimating facial surface normals `in-the-wild'. We train a fully convolutional network that can accurately recover facial normals from images including a challenging variety of expressions and facial poses. We compare against state-of-the-art face Shape-from-Shading and 3D reconstruction techniques and show that the proposed network can recover substantially more accurate and realistic normals. Furthermore, in contrast to other existing face-specific surface recovery methods, we do not require the solving of an explicit alignment step due to the fully convolutional nature of our network.
**********

3D Face Morphable Models "In-The-Wild"
James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 48-57
3D Morphable Models (3DMMs) are powerful statistical models of 3D facial shape and texture, and among the state-of-the-art methods for reconstructing facial shape from single images. With the advent of new 3D sensors, many 3D facial datasets have been collected containing both neutral as well as expressive faces. However, all datasets are captured under controlled conditions. Thus, even though powerful 3D facial shape models can be learnt from such data, it is difficult to build statistical texture models that are sufficient to reconstruct faces captured in unconstrained conditions ("in-the-wild"). In this paper, we propose the first, to the best of our knowledge, "in-the-wild" 3DMM by combining a powerful statistical model of facial shape, which describes both identity and expression, with an "in-the-wild" texture model. We show that the employment of such an "in-the-wild" texture model greatly simplifies the fitting procedure, because there is no need to optimise with regards to the illumination parameters. Furthermore, we propose a new fast algorithm for fitting the 3DMM in arbitrary images. Finally, we have captured the first 3D facial database with relatively unconstrained conditions and report quantitative evaluations with state-of-the-art performance. Complementary qualitative reconstruction results are demonstrated on standard "in-the-wild" facial databases.
**********

Towards a Quality Metric for Dense Light Fields
Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafal K. Mantiuk, Karol Myszk

owski, Hans-Peter Seidel, Piotr Didyk; Proceedings of the IEEE Conference on Com
puter Vision and Pattern Recognition (CVPR), 2017, pp. 58-67
Light fields become a popular representation of three-dimensional scenes, and th
ere is interest in their processing, resampling, and compression. As those opera
tions often result in loss of quality, there is a need to quantify it. In this w
ork, we collect a new dataset of dense reference and distorted light fields as w
ell as the corresponding quality scores which are scaled in perceptual units. Th
e scores were acquired in a subjective experiment using an interactive light-fie
ld viewing setup. The dataset contains typical artifacts that occur in light-fie
ld processing chain due to light-field reconstruction, multi-view compression, a
nd limitations of automultiscopic displays. We test a number of existing objecti
ve quality metrics to determine how well they can predict the quality of light f
ields. We find that the existing image quality metrics provide good measures of
light-field quality, but require dense reference light- fields for optimal perfo
rmance. For more complex tasks of comparing two distorted light fields, their pe
rformance drops significantly, which reveals the need for new, light-field-speci
fic metrics.
**********

Position Tracking for Virtual Reality Using Commodity WiFi
Manikanta Kotaru, Sachin Katti; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2017, pp. 68-78
Today, experiencing virtual reality (VR) is a cumbersome experience which either
 requires dedicated infrastructure like infrared cameras to track the headset an
d hand-motion controllers (e.g., Oculus Rift, HTC Vive), or provides only 3-DoF
(Degrees of Freedom) tracking which severely limits the user experience (e.g., S
amsung Gear). To truly enable VR everywhere, we need position tracking to be ava
ilable as a ubiquitous service. This paper presents WiCapture, a novel approach
which leverages commodity WiFi infrastructure, which is ubiquitous today, for tr
acking purposes. We prototype WiCapture using off-the-shelf WiFi radios and show
 that it achieves an accuracy of 0.88 cm compared to sophisticated infrared-base
d tracking systems like the Oculus, while providing much higher range, resistanc
e to occlusion, ubiquity and ease of deployment.
**********

Material Classification Using Frequency- and Depth-Dependent Time-Of-Flight Dist
ortion
Kenichiro Tanaka, Yasuhiro Mukaigawa, Takuya Funatomi, Hiroyuki Kubo, Yasuyuki M
atsushita, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2017, pp. 79-88
This paper presents a material classification method using an off-the-shelf Time
-of-Flight (ToF) camera. We use a key observation that the depth measurement by
a ToF camera is distorted in objects with certain materials, especially with tra
nslucent materials. We show that this distortion is caused by the variations of
time domain impulse responses across materials and also by the measurement mecha
nism of the existing ToF cameras. Specifically, we reveal that the amount of dis
tortion varies according to the modulation frequency of the ToF camera, the mate
rial of the object, and the distance between the camera and object. Our method u
ses the depth distortion of ToF measurements as features and achieves material c
lassification of a scene. Effectiveness of the proposed method is demonstrated b
y numerical evaluation and real-world experiments, showing its capability of eve
n classifying visually similar objects.
**********

Learning by Association -- A Versatile Semi-Supervised Training Method for Neura
l Networks
Philip Haeusser, Alexander Mordvintsev, Daniel Cremers; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 89-98
In many real-world scenarios, labeled data for a specific machine learning task
is costly to obtain. Semi-supervised training methods make use of abundantly ava
ilable unlabeled data and a smaller number of labeled examples. We propose a new
 framework for semi-supervised training of deep neural networks inspired by lear
ning in humans. "Associations" are made from embeddings of labeled samples to th

ose of unlabeled ones and back. The optimization schedule encourages correct ass
ociation cycles that end up at the same class from which the association was sta
rted and penalizes wrong associations ending at a different class. The implement
ation is easy to use and can be added to any existing end-to-end training setup.
 We demonstrate the capabilities of learning by association on several data sets
 and show that it can improve performance on classification tasks tremendously b
y making use of additionally available unlabeled data. In particular, for cases
with few labeled data, our training scheme outperforms the current state of the
art on SVHN.
**********

A Non-Convex Variational Approach to Photometric Stereo Under Inaccurate Lightin
g

Yvain Queau, Tao Wu, Francois Lauze, Jean-Denis Durou, Daniel Cremers; Proceedin
gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201
7, pp. 99-108

This paper tackles the photometric stereo problem in the presence of inaccurate
lighting, obtained either by calibration or by an uncalibrated photometric stere
o method. Based on a precise modeling of noise and outliers, a robust variationa
l approach is introduced. It explicitly accounts for self-shadows, and enforces
robustness to cast-shadows and specularities by resorting to redescending M-esti
mators. The resulting non-convex model is solved by means of a computationally e
fficient alternating reweighted least-squares algorithm. Since it implicitly enf
orces integrability, the new variational approach can refine both the intensitie
s and the directions of the lighting.
**********

Learning From Synthetic Humans

Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan
 Laptev, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition (CVPR), 2017, pp. 109-117

Estimating human pose, shape, and motion from images and video are fundamental c
hallenges with many applications. Recent advances in 2D human pose estimation us
e large amounts of manually-labeled training data for learning convolutional neu
ral networks (CNNs). Such data is time consuming to acquire and difficult to ext
end. Moreover, manual labeling of 3D pose, depth and motion is impractical. In t
his work we present SURREAL: a new large-scale dataset with synthetically-genera
ted but realistic images of people rendered from 3D sequences of human motion ca
pture data. We generate more than 6 million frames together with ground truth po
se, depth maps, and segmentation masks. We show that CNNs trained on our synthet
ic dataset allow for accurate human depth estimation and human part segmentation
 in real RGB images. Our results and the new datast open up new possibilities fo
r advancing person analysis using chap and large-scale synthetic data.
**********

Correlational Gaussian Processes for Cross-Domain Visual Recognition

Chengjiang Long, Gang Hua; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2017, pp. 118-126

We present a probabilistic model that captures higher order co-occurrence statis
tics for joint visual recognition in a collection of images and across multiple
domains. More importantly, we predict the structured output across multiple doma
ins by correlating outputs from the multi-classes Gaussian process classifiers i
n each individual domain. A set of correlational tensors is adopted to model the
 relationship within a single domain as well as across multiple domains. This re
nders it possible to explore a high-order relational model instead of using just
 a set of pairwise relational models. Such tensor relations are based on both th
e positive and negative co-occurrences of different categories of visual instanc
es across multi-domains. This is in contrast to most previous models where only
pair-wise relationships are explored. We conduct experiments on four challenging
 image collections. The experimental results clearly demonstrate the efficacy of
 our proposed model.
**********

Revisiting the Variable Projection Method for Separable Nonlinear Least Squares

Problems
Je Hyeong Hong, Christopher Zach, Andrew Fitzgibbon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 127-135
Variable Projection (VarPro) is a framework to solve optimization problems efficiently by optimally eliminating a subset of the unknowns. It is in particular adapted for Separable Nonlinear Least Squares (SNLS) problems, a class of optimization problems including low-rank matrix factorization with missing data and affine bundle adjustment as instances. VarPro-based methods have received much attention over the last decade due to the experimentally observed large convergence basin for certain problem classes, where they have a clear advantage over standard methods based on Joint optimization over all unknowns. Yet no clear answers have been found in the literature as to why VarPro outperforms others and why Joint optimization, which has been successful in solving many computer vision tasks, fails on this type of problems. Also, the fact that VarPro has been mainly tested on small to medium-sized datasets has raised questions about its scalability. This paper intends to address these unsolved puzzles.
**********

Learning to Detect Salient Objects With Image-Level Supervision
Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, Xiang Ruan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 136-145
Deep Neural Networks (DNNs) have substantially improved the state-of-the-art in salient object detection. However, training DNNs requires costly pixel-level annotations. In this paper, we leverage the observation that image-level tags provide important cues of foreground salient objects, and develop a weakly supervised learning method for saliency detection using image-level tags only. The Foreground Inference Network (FIN) is introduced for this challenging task. In the first stage of our training method, FIN is jointly trained with a fully convolutional network (FCN) for image-level tag prediction. A global smooth pooling layer is proposed, enabling FCN to assign object category tags to corresponding object regions, while FIN is capable of capturing all potential foreground regions with the predicted saliency maps. In the second stage, FIN is fine-tuned with its predicted saliency maps as ground truth. For refinement of ground truth, an iterative Conditional Random Field is developed to enforce spatial label consistency and further boost performance. Our method alleviates annotation efforts and allows the usage of existing large scale training sets with image-level tags. Our model runs at 60 FPS, outperforms unsupervised ones with a large margin, and achieves comparable or even superior performance than fully supervised counterparts.
**********

Binary Coding for Partial Action Analysis With Limited Observation Ratios
Jie Qin, Li Liu, Ling Shao, Bingbing Ni, Chen Chen, Fumin Shen, Yunhong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 146-155
Traditional action recognition methods aim to recognize actions with complete observations/executions. However, it is often difficult to capture fully executed actions due to occlusions, interruptions, etc. Meanwhile, action prediction/recognition in advance based on partial observations is essential for preventing the situation from deteriorating. Besides, fast spotting human activities using partially observed data is a critical ingredient for retrieval systems. Inspired by the recent success of data binarization in efficient retrieval/recognition, we propose a novel approach, named Partial Reconstructive Binary Coding (PRBC), for action analysis based on limited frame glimpses during any period of the complete execution. Specifically, we learn discriminative compact binary codes for partial actions via a joint learning framework, which collaboratively tackles feature reconstruction as well as binary coding. We obtain the solution to PRBC based on a discrete alternating iteration algorithm. Extensive experiments on four realistic action datasets in terms of three tasks (i.e., partial action retrieval, recognition and prediction) clearly show the superiority of PRBC over the state-of-the-art methods, along with significantly reduced memory load and computational costs during the online test.

**********
Temporal Convolutional Networks for Action Segmentation and Detection
Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, Gregory D. Hager; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 156-165
The ability to identify and temporally segment fine-grained human actions throughout a video is crucial for robotics, surveillance, education, and beyond. Typical approaches decouple this problem by first extracting local spatiotemporal features from video frames and then feeding them into a temporal classifier that captures high-level temporal patterns. We describe a class of temporal models, which we call Temporal Convolutional Networks (TCNs), that use a hierarchy of temporal convolutions to perform fine-grained action segmentation or detection. Our Encoder-Decoder TCN uses pooling and upsampling to efficiently capture long-range temporal patterns whereas our Dilated TCN uses dilated convolutions. We show that TCNs are capable of capturing action compositions, segment durations, and long-range dependencies, and are over a magnitude faster to train than competing LSTM-based Recurrent Neural Networks. We apply these models to three challenging fine-grained datasets and show large improvements over the state of the art.
**********
DeLiGAN : Generative Adversarial Networks for Diverse and Limited Data
Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, R. Venkatesh Babu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 166-174
A class of recent approaches for generating images, called Generative Adversarial Networks (GAN), have been used to generate impressively realistic images of objects, bedrooms, handwritten digits and a variety of other image modalities. However, typical GAN-based approaches require large amounts of training data to capture the diversity across the image modality. In this paper, we propose DeLiGAN -- a novel GAN-based architecture for diverse and limited training data scenarios. In our approach, we reparameterize the latent generative space as a mixture model and learn the mixture model's parameters along with those of GAN. This seemingly simple modification to the GAN framework is surprisingly effective and results in models which enable diversity in generated samples although trained with limited data. In our work, we show that DeLiGAN can generate images of handwritten digits, objects and hand-drawn sketches, all using limited amounts of data. To quantitatively characterize intra-class diversity of generated samples, we also introduce a modified version of "inception-score", a measure which has been found to correlate well with human assessment of generated samples.
**********
Template Matching With Deformable Diversity Similarity
Itamar Talmi, Roey Mechrez, Lihi Zelnik-Manor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 175-183
We propose a novel measure for template matching named Deformable Diversity Similarity -- based on the diversity of feature matches between a target image window and the template. We rely on both local appearance and geometric information that jointly lead to a powerful approach for matching. Our key contribution is a similarity measure, that is robust to complex deformations, significant background clutter, and occlusions. Empirical evaluation on the most up-to-date benchmark shows that our method outperforms the current state-of-the-art in its detection accuracy while improving computational complexity.
**********
Surface Motion Capture Transfer With Gaussian Process Regression
Adnane Boukhayma, Jean-Sebastien Franco, Edmond Boyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 184-192
We address the problem of transferring motion between captured 4D models. We particularly focus on human subjects for which the ability to automatically augment 4D datasets, by propagating movements between subjects, is of interest in a great deal of recent vision applications that builds on human visual corpus. Given 4D training sets for two subjects for which a sparse set of corresponding keyposes are known, our method is able to transfer a newly captured motion from one su

bject to the other. With the aim to generalize transfers to input motions possib ly very diverse with respect to the training sets, the method contributes with a new transfer model based on non-linear pose interpolation. Building on Gaussian process regression, this model intends to capture and preserve individual motio n properties, and thereby realism, by accounting for pose inter-dependencies dur ing motion transfers. Our experiments show visually qualitative, and quantitativ e, improvements over existing pose-mapping methods and confirm the generalizatio n capabilities of our method compared to state of the art.
**********

Generating Holistic 3D Scene Abstractions for Text-Based Image Retrieval
Ang Li, Jin Sun, Joe Yue-Hei Ng, Ruichi Yu, Vlad I. Morariu, Larry S. Davis; Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ), 2017, pp. 193-201
Spatial relationships between objects provide important information for text-bas ed image retrieval. As users are more likely to describe a scene from a real wor ld perspective, using 3D spatial relationships rather than 2D relationships that assume a particular viewing direction, one of the main challenges is to infer t he 3D structure that bridges images with users' text descriptions. However, dire ct inference of 3D structure from images requires learning from large scale anno tated data. Since interactions between objects can be reduced to a limited set o f atomic spatial relations in 3D, we study the possibility of inferring 3D struc ture from a text description rather than an image, applying physical relation mo dels to synthesize holistic 3D abstract object layouts satisfying the spatial co nstraints present in a textual description. We present a generic framework for r etrieving images from a textual description of a scene by matching images with t hese generated abstract object layouts. Images are ranked by matching object det ection outputs (bounding boxes) to 2D layout candidates (also represented by bou nding boxes) which are obtained by projecting the 3D scenes with sampled camera directions. We validate our approach using public indoor scene datasets and show that our method outperforms baselines built upon object occurrence histograms a nd learned 2D pairwise relations.
**********

Unsupervised Video Summarization With Adversarial LSTM Networks
Behrooz Mahasseni, Michael Lam, Sinisa Todorovic; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 202-211
This paper addresses the problem of unsupervised video summarization, formulated as selecting a sparse subset of video frames that optimally represent the input video. Our key idea is to learn a deep summarizer network to minimize distance between training videos and a distribution of their summarizations, in an unsupe rvised way. Such a summarizer can then be applied on a new video for estimating its optimal summarization. For learning, we specify a novel generative adversari al framework, consisting of the summarizer and discriminator. The summarizer is the autoencoder long short-term memory network (LSTM) aimed at, first, selecting video frames, and then decoding the obtained summarization for reconstructing t he input video. The discriminator is another LSTM aimed at distinguishing betwee n the original video and its reconstruction from the summarizer. The summarizer LSTM is cast as an adversary of the discriminator, i.e., trained so as to maxima lly confuse the discriminator. This learning is also regularized for sparsity. E valuation on four benchmark datasets, consisting of videos showing diverse event s in first- and third-person views, demonstrates our competitive performance in comparison to fully supervised state-of-the-art approaches.
**********

SphereFace: Deep Hypersphere Embedding for Face Recognition
Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, Le Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 212-220
This paper addresses deep face recognition (FR) problem under open-set protocol, where ideal face features are expected to have smaller maximal intra-class dist ance than minimal inter-class distance under a suitably chosen metric space. How ever, few existing algorithms can effectively achieve this criterion. To this en

d, we propose the angular softmax (A-Softmax) loss that enables convolutional ne
ural networks (CNNs) to learn angularly discriminative features. Geometrically,
A-Softmax loss can be viewed as imposing discriminative constraints on a hypersp
here manifold, which intrinsically matches the prior that faces also lie on a ma
nifold. Moreover, the size of angular margin can be quantitatively adjusted by a
 parameter m. We further derive specific m to approximate the ideal feature crit
erion. Extensive analysis and experiments on Labeled Face in the Wild (LFW), You
tube Faces (YTF) and MegaFace Challenge 1 show the superiority of A-Softmax loss
 in FR tasks.
**********

One-Shot Video Object Segmentation
Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixe, Daniel
 Cremers, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision an
d Pattern Recognition (CVPR), 2017, pp. 221-230
This paper tackles the task of semi-supervised video object segmentation, i.e.,
the separation of an object from the background in a video, given the mask of th
e first frame. We present One-Shot Video Object Segmentation (OSVOS), based on a
 fully-convolutional neural network architecture that is able to successively tr
ansfer generic semantic information, learned on ImageNet, to the task of foregro
und segmentation, and finally to learning the appearance of a single annotated o
bject of the test sequence (hence one-shot). Although all frames are processed i
ndependently, the results are temporally coherent and stable. We perform experim
ents on two annotated video segmentation databases, which show that OSVOS is fas
t and improves the state of the art by a significant margin (79.8% vs 68.0%).
**********

SGM-Nets: Semi-Global Matching With Neural Networks
Akihito Seki, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2017, pp. 231-240
This paper deals with deep neural networks for predicting accurate dense dispari
ty map with Semi-global matching (SGM). SGM is a widely used regularization meth
od for real scenes because of its high accuracy and fast computation speed. Even
 though SGM can obtain accurate results, tuning of SGM's penalty-parameters, whi
ch control a smoothness and discontinuity of a disparity map, is uneasy and empi
rical methods have been proposed. We propose a learning based penalties estimati
on method, which we call SGM-Nets that consist of Convolutional Neural Networks.
 A small image patch and its position are input into SGMNets to predict the pena
lties for the 3D object structures. In order to train the networks, we introduce
 a novel loss function which is able to use sparsely annotated disparity maps su
ch as captured by a LiDAR sensor in real environments. Moreover, we propose a no
vel SGM parameterization, which deploys different penalties depending on either
positive or negative disparity changes in order to represent the object structur
es more discriminatively. Our SGM-Nets outperformed state of the art accuracy on
 KITTI benchmark datasets.
**********

What's in a Question: Using Visual Questions as a Form of Supervision
Siddha Ganju, Olga Russakovsky, Abhinav Gupta; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 241-250
Collecting fully annotated image datasets is challenging and expensive. Many typ
es of weak supervision have been explored: weak manual annotations, web search r
esults, temporal continuity, ambient sound and others. We focus on one particula
r unexplored mode: visual questions that are asked about images. The key observa
tion that inspires our work is that the question itself provides useful informat
ion about the image (even without the answer being available). For instance, the
 question "what is the breed of the dog?" informs the AI that the animal in the
scene is a dog and that there is only one dog present. We make three contributio
ns: (1) providing an extensive qualitative and quantitative analysis of the info
rmation contained in human visual questions, (2) proposing two simple but surpri
singly effective modifications to the standard visual question answering models
that allow them to make use of weak supervision in the form of unanswered questi
ons associated with images and (3) demonstrating that a simple data augmentation

strategy inspired by our insights results in a 7.1% improvement on the standard VQA benchmark.
**********

Context-Aware Captions From Context-Agnostic Supervision
Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, Gal Chechik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 251-260
We introduce an inference technique to produce discriminative context-aware image captions (captions that describe differences between images or visual concepts) using only generic context-agnostic training data (captions that describe a concept or an image in isolation). For example, given images and captions of "siamese cat" and "tiger cat", we generate language that describes the "siamese cat" in a way that distinguishes it from "tiger cat". Our key novelty is that we show how to do joint inference over a language model that is context-agnostic and a listener which distinguishes closely-related concepts. We first apply our technique to a justification task, namely to describe why an image contains a particular fine-grained category as opposed to another closely-related category of the CUB- 200-2011 dataset. We then study discriminative image captioning to generate language that uniquely refers to one of two semantically-similar images in the COCO dataset. Evaluations with discriminative ground truth for justification and human studies for discriminative image captioning reveal that our approach outperforms baseline generative and speaker-listener approaches for discrimination.
**********

Polyhedral Conic Classifiers for Visual Object Detection and Classification
Hakan Cevikalp, Bill Triggs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 261-269
We propose a family of quasi-linear discriminants that outperform current large-margin methods in sliding window visual object detection and open set recognition tasks. In these tasks the classification problems are both numerically imbalanced -- positive (object class) training and test windows are much rarer than negative (non-class) ones -- and geometrically asymmetric -- the positive samples typically form compact, visually-coherent groups while negatives are much more diverse, including anything at all that is not a well-centred sample from the target class. It is difficult to cover such negative classes using training samples, and doubly so in `open set' applications where run-time negatives may stem from classes that were not seen at all during training. So there is a need for discriminants whose decision regions focus on tightly circumscribing the positive class, while still taking account of negatives in zones where the two classes overlap. This paper introduces a family of quasi-linear "polyhedral conic" discriminants whose positive regions are distorted L1 balls. The methods have properties and run-time complexities comparable to linear Support Vector Machines (SVMs), and they can be trained from either binary or positive-only samples using constrained quadratic programs related to SVMs. Our experiments show that they significantly outperform both linear SVMs and existing one-class discriminants on a wide range of object detection, open set recognition and conventional closed-set classification tasks.
**********

Unsupervised Monocular Depth Estimation With Left-Right Consistency
Clement Godard, Oisin Mac Aodha, Gabriel J. Brostow; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 270-279
Learning based methods have shown very promising results for the task of depth estimation in single images. However, most existing approaches treat depth prediction as a supervised regression problem and as a result, require vast quantities of corresponding ground truth depth data for training. Just recording quality depth data in a range of environments is a challenging problem. In this paper, we innovate beyond existing approaches, replacing the use of explicit depth data during training with easier-to-obtain binocular stereo footage. We propose a novel training objective that enables our convo- lutional neural network to learn to perform single image depth estimation, despite the absence of ground truth depth data. Ex- ploiting epipolar geometry constraints, we generate disparity images

by training our network with an image reconstruction loss. We show that solving for image reconstruction alone re- sults in poor quality depth images. To overcome this problem, we propose a novel training loss that enforces consistency be- tween the disparities produced relative to both the left and right images, leading to improved performance and robustness com- pared to existing approaches. Our method produces state of the art results for monocular depth estimation on the KITTI driving dataset, even outperforming supervised methods that have been trained with ground truth depth.

**********

Compact Matrix Factorization With Dependent Subspaces

Viktor Larsson, Carl Olsson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 280-289

Traditional matrix factorization methods approximate high dimensional data with a low dimensional subspace. This imposes constraints on the matrix elements which allow for estimation of missing entries. A lower rank provides stronger constraints and makes estimation of the missing entries less ambiguous at the cost of measurement fit. In this paper we propose a new factorization model that further constrains the matrix entries. Our approach can be seen as a unification of traditional low-rank matrix factorization and the more recent union-of-subspace approach. It adaptively finds clusters that can be modeled with low dimensional local subspaces and simultaneously uses a global rank constraint to capture the overall scene interactions. For inference we use an energy that penalizes a trade-off between data fit and degrees-of-freedom of the resulting factorization. We show qualitatively and quantitatively that regularizing both local and global dynamics yields significantly improved missing data estimation.

**********

Deep Reinforcement Learning-Based Image Captioning With Embedding Reward

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, Li-Jia Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 290-298

Image captioning is a challenging problem owing to the complexity in understanding the image content and diverse ways of describing it in natural language. Recent advances in deep neural networks have substantially improved the performance of this task. Most state-of-the-art approaches follow an encoder-decoder framework, which generates captions using a sequential recurrent prediction model. However, in this paper, we introduce a novel decision-making framework for image captioning. We utilize a "policy network" and a "value network" to collaboratively generate captions. The policy network serves as a local guidance by providing the confidence of predicting the next word according to the current state. Additionally, the value network serves as a global and lookahead guidance by evaluating all possible extensions of the current state. In essence, it adjusts the goal of predicting the correct words towards the goal of generating captions similar to the ground truth captions. We train both networks using an actor-critic reinforcement learning model, with a novel reward defined by visual-semantic embedding. Extensive experiments and analyses on the Microsoft COCO dataset show that the proposed framework outperforms state-of-the-art approaches across different evaluation metrics.

**********

Dual Attention Networks for Multimodal Reasoning and Matching

Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 299-307

We propose Dual Attention Networks (DANs) which jointly leverage visual and textual attention mechanisms to capture fine-grained interplay between vision and language. DANs attend to specific regions in images and words in text through multiple steps and gather essential information from both modalities. Based on this framework, we introduce two types of DANs for multimodal reasoning and matching, respectively. The reasoning model allows visual and textual attentions to steer each other during collaborative inference, which is useful for tasks such as Visual Question Answering (VQA). In addition, the matching model exploits the two attention mechanisms to estimate the similarity between images and sentences by focusing on their shared semantics. Our extensive experiments validate the effec

tiveness of DANs in combining vision and language, achieving the state-of-the-art performance on public benchmarks for VQA and image-text matching.
**********

Exploiting 2D Floorplan for Building-Scale Panorama RGBD Alignment
Erik Wijmans, Yasutaka Furukawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 308-316
This paper presents a novel algorithm that utilizes a 2D floorplan to align panorama RGBD scans. While effective panorama RGBD alignment techniques exist, such a system requires extremely dense RGBD image sampling. Our approach can significantly reduce the number of necessary scans with the aid of a floorplan image. We formulate a novel Markov Random Field inference problem as a scan placement over the floorplan, as opposed to the conventional scan-to-scan alignment. The technical contributions lie in multi-modal image correspondence cues (between scans and schematic floorplan) as well as a novel coverage potential avoiding an inherent stacking bias. The proposed approach has been evaluated on five challenging large indoor spaces. To the best of our knowledge, we present the first effective system that utilizes a 2D floorplan image for building-scale 3D pointcloud alignment. The source code and the data are shared with the community to further enhance indoor mapping research.
**********

A Hierarchical Approach for Generating Descriptive Image Paragraphs
Jonathan Krause, Justin Johnson, Ranjay Krishna, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 317-325
Recent progress on image captioning has made it possible to generate novel sentences describing images in natural language, but compressing an image into a single sentence can describe visual content in only coarse detail. While one new captioning approach, dense captioning, can potentially describe images in finer levels of detail by captioning many regions within an image, it in turn is unable to produce a coherent story for an image. In this paper we overcome these limitations by generating entire paragraphs for describing images, which can tell detailed, unified stories. We develop a model that decomposes both images and paragraphs into their constituent parts, detecting semantic regions in images and using a hierarchical recurrent neural network to reason about language. Linguistic analysis confirms the complexity of the paragraph generation task, and thorough experiments on a new dataset of image and paragraph pairs demonstrate the effectiveness of our approach.
**********

Visual Dialog
Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 326-335
We introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a question about the image, the agent has to ground the question in image, infer context from history, and answer the question accurately. Visual Dialog is disentangled enough from a specific downstream task so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress. We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). VisDial contains 1 dialog (10 question- answer pairs) on  140k images from the COCO dataset, with a total of  1.4M dialog question-answer pairs. We introduce a family of neural encoder-decoder models for Visual Dialog with 3 encoders (Late Fusion, Hierarchical Recurrent Encoder and Memory Network) and 2 decoders (generative and discriminative), which outperform a number of sophisticated baselines. We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a set of candidate answers and evaluated on metrics such as mean-reciprocal-rank of human response. We quantify gap between machine and human performance on the Visual Dialog task via human studies. Our dataset, code, an

d trained models will be released publicly at https://visualdialog.org. Putting it all together, we demonstrate the first 'visual chatbot'!
**********

DESIRE: Distant Future Prediction in Dynamic Scenes With Interacting Agents
Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 336-345
We introduce a Deep Stochastic IOC RNN Encoder-decoder framework, DESIRE, for the task of future predictions of multiple interacting agents in dynamic scenes. DESIRE effectively predicts future locations of objects in multiple scenes by 1) accounting for the multi-modal nature of the future prediction (i.e., given the same context, future may vary), 2) foreseeing the potential future outcomes and make a strategic prediction based on that, and 3) reasoning not only from the past motion history, but also from the scene context as well as the interactions among the agents. DESIRE achieves these in a single end-to-end trainable neural network model, while being computationally efficient. The model first obtains a diverse set of hypothetical future prediction samples employing a conditional variational auto-encoder, which are ranked and refined by the following RNN scoring-regression module. Samples are scored by accounting for accumulated future rewards, which enables better long-term strategic decisions similar to IOC frameworks. An RNN scene context fusion module jointly captures past motion histories, the semantic scene context and interactions among multiple agents. A feedback mechanism iterates over the ranking and refinement to further boost the prediction accuracy. We evaluate our model on two publicly available datasets: KITTI and Stanford Drone Dataset. Our experiments show that the proposed model significantly improves the prediction accuracy compared to other baseline methods.
**********

Mining Object Parts From CNNs via Active Question-Answering
Quanshi Zhang, Ruiming Cao, Ying Nian Wu, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 346-355
Given a convolutional neural network (CNN) that is pre-trained for object classification, this paper proposes to use active question-answering to semanticize neural patterns in conv-layers of the CNN and mine part concepts. For each part concept, we mine neural patterns in the pre-trained CNN, which are related to the target part, and use these patterns to construct an And-Or graph (AOG) to represent a four-layer semantic hierarchy of the part. As an interpretable model, the AOG associates different CNN units with different explicit object parts. We use an active human-computer communication to incrementally grow such an AOG on the pre-trained CNN as follows. We allow the computer to actively identify objects, whose neural patterns cannot be explained by the current AOG. Then, the computer asks human about the unexplained objects, and uses the answers to automatically discover certain CNN patterns corresponding to the missing knowledge. We incrementally grow the AOG to encode new knowledge discovered during the active-learning process. In experiments, our method exhibits high learning efficiency. Our method uses about 1/6--1/3 of the part annotations for training, but achieves similar or better part-localization performance than fast-RCNN methods.
**********

Multi-Way Multi-Level Kernel Modeling for Neuroimaging Classification
Lifang He, Chun-Ta Lu, Hao Ding, Shen Wang, Linlin Shen, Philip S. Yu, Ann B. Ragin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 356-364
Owing to prominence as a diagnostic tool for probing the neural correlates of cognition, neuroimaging tensor data has been the focus of intense investigation. Although many supervised tensor learning approaches have been proposed, they either cannot capture the nonlinear relationships of tensor data or cannot preserve the complex multi-way structural information. In this paper, we propose a Multi-way Multi-level Kernel (MMK) model that can extract discriminative, nonlinear and structural preserving representations of tensor data. Specifically, we introduce a kernelized CP tensor factorization technique, which is equivalent to performing the low-rank tensor factorization in a possibly much higher dimensional spa

ce that is implicitly defined by the kernel function. We further employ a multi-way nonlinear feature mapping to derive the dual structural preserving kernels, which are used in conjunction with kernel machines (e.g., SVM). Extensive experiments on real-world neuroimages demonstrate that the proposed MMK method can effectively boost the classification performance on diverse brain disorders (i.e., Alzheimer's disease, ADHD, and HIV).
**********

Low-Rank Bilinear Pooling for Fine-Grained Classification
Shu Kong, Charless Fowlkes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 365-374
Pooling second-order local feature statistics to form a high-dimensional bilinear feature has been shown to achieve state-of-the-art performance on a variety of fine-grained classification tasks. To address the computational demands of high feature dimensionality, we propose to represent the covariance features as a matrix and apply a low-rank bilinear classifier. The resulting classifier can be evaluated without explicitly computing the bilinear feature map which allows for a large reduction in the compute time as well as decreasing the effective number of parameters to be learned. To further compress the model, we propose a classifier co-decomposition that factorizes the collection of bilinear classifiers into a common factor and compact per-class terms. The co-decomposition idea can be deployed through two convolutional layers and trained in an end-to-end architecture. We suggest a simple yet effective initialization that avoids explicitly first training and factorizing the larger bilinear classifiers. Through extensive experiments, we show that our model achieves state-of-the-art performance on several public datasets for fine-grained classification trained with only category labels. Importantly, our final model is an order of magnitude smaller than the recently proposed compact bilinear model [??], and three orders smaller than the standard bilinear CNN model [??].
**********

Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning
Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 375-383
Attention-based neural encoder-decoder frameworks have been widely adopted for image captioning. Most methods force visual attention to be active for every generated word. However, the decoder likely requires little to no visual information from the image to predict non-visual words such as "the" and "of". Other words that may seem visual can often be predicted reliably just from the language model e.g., "sign" after "behind a red stop" or "phone" following "talking on a cell". In this paper, we propose a novel adaptive attention model with a visual sentinel. At each time step, our model decides whether to attend to the image (and if so, to which regions) or to the visual sentinel. The model decides whether to attend to the image and where, in order to extract meaningful information for sequential word generation. We test our method on the COCO image captioning 2015 challenge dataset and Flickr30K. Our approach sets the new state-of-the-art by a significant margin.
**********

Learning Deep Context-Aware Features Over Body and Latent Parts for Person Re-Identification
Dangwei Li, Xiaotang Chen, Zhang Zhang, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 384-393
Person Re-identification (ReID) is to identify the same person across different cameras. It is a challenging task due to the large variations in person pose, occlusion, background clutter, etc. How to extract powerful features is a fundamental problem in ReID and is still an open problem today. In this paper, we design a Multi-Scale Context-Aware Network (MSCAN) to learn powerful features over full body and body parts, which can well capture the local context knowledge by stacking multi-scale convolutions in each layer. Moreover, instead of using predefined rigid parts, we propose to learn and localize deformable pedestrian parts using Spatial Transformer Networks (STN) with novel spatial constraints. The learn

ed body parts can release some difficulties, e.g. pose variations and background clutters, in part-based representation. Finally, we integrate the representation learning processes of full body and body parts into a unified framework for person ReID through multi-class person identification tasks. Extensive evaluations on current challenging large-scale person ReID datasets, including the image-based Market1501, CUHK03 and sequence-based MARS datasets, show that the proposed method achieves the state-of-the-art results.
**********

Turning an Urban Scene Video Into a Cinemagraph
Hang Yan, Yebin Liu, Yasutaka Furukawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 394-402
This paper proposes an algorithm that turns a regular video capturing urban scenes into a high-quality endless animation, known as a Cinemagraph. The creation of a Cinemagraph usually requires a static camera in a carefully configured scene. The task becomes challenging for a regular video with a moving camera and objects. Our approach first warps an input video into the viewpoint of a reference camera. Based on the warped video, we propose effective temporal analysis algorithms to detect regions with static geometry and dynamic appearance, where geometric modeling is reliable and visually attractive animations can be created. Lastly, the algorithm applies a sequence of video processing techniques to produce a Cinemagraph movie. We have tested the proposed approach on numerous challenging real scenes. To our knowledge, this work is the first to automatically generate Cinemagraph animations from regular movies in the wild.
**********

Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-Identification
Weihua Chen, Xiaotang Chen, Jianguo Zhang, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 403-412
Person re-identification (ReID) is an important task in wide area video surveillance which focuses on identifying people across different cameras. Recently, deep learning networks with a triplet loss become a common framework for person ReID. However, the triplet loss pays main attentions on obtaining correct orders on the training set. It still suffers from a weaker generalization capability from the training set to the testing set, thus resulting in inferior performance. In this paper, we design a quadruplet loss, which can lead to the model output with a larger inter-class variation and a smaller intra-class variation compared to the triplet loss. As a result, our model has a better generalization ability and can achieve a higher performance on the testing set. In particular, a quadruplet deep network using a margin-based online hard negative mining is proposed based on the quadruplet loss for the person ReID. In extensive experiments, the proposed network outperforms most of the state-of-the-art algorithms on representative datasets which clearly demonstrates the effectiveness of our proposed method.
**********

Surveillance Video Parsing With Single Frame Supervision
Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, Yao Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 413-421
Surveillance video parsing, which segments the video frames into several labels, i.e., face, pants, left-leg, has wide applications. However, annotating all frames pixel-wisely is tedious and inefficient. In this paper, we develop a Single frame Video Parsing (SVP) method which requires only one labeled frame per video in training stage. To parse one particular frame, the video segment preceding the frame is jointly considered. SVP 1: roughly parses the frames within the video segment, 2: estimates the optical flow between frames and 3: fuses the rough parsing results warped by optical flow to produce the refined parsing result. The three components of SVP, namely frame parsing, optical flow estimation and temporal fusion are integrated in an end-to-end manner. Experimental results on two surveillance video datasets reveal that SVP is superior than state-of-the-arts.
**********

Semantically Coherent Co-Segmentation and Reconstruction of Dynamic Scenes
Armin Mustafa, Adrian Hilton; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2017, pp. 422-431
In this paper we propose a framework for spatially and temporally coherent seman tic co-segmentation and reconstruction of complex dynamic scenes from multiple s tatic or moving cameras. Semantic co-segmentation exploits the coherence in sema ntic class labels both spatially, between views at a single time instant, and te mporally, between widely spaced time instants of dynamic objects with similar sh ape and appearance. We demonstrate that semantic coherence results in improved s egmentation and reconstruction for complex scenes. A joint formulation is propos ed for semantically coherent object-based co-segmentation and reconstruction of scenes by enforcing consistent semantic labelling between views and over time. S emantic tracklets are introduced to enforce temporal coherence in semantic label ling and reconstruction between widely spaced instances of dynamic objects. Trac klets of dynamic objects enable unsupervised learning of appearance and shape pr iors that are exploited in joint segmentation and reconstruction. Evaluation on challenging indoor and outdoor sequences with hand-held moving cameras shows imp roved accuracy in segmentation, temporally coherent semantic labelling and 3D re construction of dynamic scenes.
**********

Transition Forests: Learning Discriminative Temporal Transitions for Action Reco gnition and Detection
Guillermo Garcia-Hernando, Tae-Kyun Kim; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2017, pp. 432-440
A human action can be seen as transitions between one's body poses over time, wh ere the transition depicts a temporal relation between two poses. Recognizing ac tions thus involves learning a classifier sensitive to these pose transitions as well as to static poses. In this paper, we introduce a novel method called tran sitions forests, an ensemble of decision trees that both learn to discriminate s tatic poses and transitions between pairs of two independent frames. During trai ning, node splitting is driven by alternating two criteria: the standard classif ication objective that maximizes the discrimination power in individual frames, and the proposed one in pairwise frame transitions. Growing the trees tends to g roup frames that have similar associated transitions and share same action label incorporating temporal information that was not available otherwise. Unlike con ventional decision trees where the best split in a node is determined independen tly of other nodes, the transition forests try to find the best split of nodes j ointly (within a layer) for incorporating distant node transitions. When inferri ng the class label of a new frame, it is passed down the trees and the predictio n is made based on previous frame predictions and the current one in an efficien t and online manner. We apply our method on varied skeleton action recognition a nd online detection datasets showing its suitability over several baselines and state-of-the-art approaches.
**********

Pixelwise Instance Segmentation With a Dynamically Instantiated Network
Anurag Arnab, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 441-450
Semantic segmentation and object detection research have recently achieved rapid progress. However, the former task has no notion of different instances of the same object, and the latter operates at a coarse, bounding-box level. We propose an Instance Segmentation system that produces a segmentation map where each pix el is assigned an object class and instance identity label. Most approaches adap t object detectors to produce segments instead of boxes. In contrast, our method is based on an initial semantic segmentation module, which feeds into an instan ce subnetwork. This subnetwork uses the initial category-level segmentation, alo ng with cues from the output of an object detector, within an end-to-end CRF to predict instances. This part of our model is dynamically instantiated to produce a variable number of instances per image. Our end-to-end approach requires no p ost-processing and considers the image holistically, instead of processing indep endent proposals. Therefore, unlike some related work, a pixel cannot belong to

multiple instances. Furthermore, far more precise segmentations are achieved, as shown by our state-of-the-art results (particularly at high IoU thresholds) on the Pascal VOC and Cityscapes datasets.
**********

Video Propagation Networks
Varun Jampani, Raghudeep Gadde, Peter V. Gehler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 451-461
We propose a technique that propagates information forward through video data. The method is conceptually simple and can be applied to tasks that require the propagation of structured information, such as semantic labels, based on video content. We propose a "Video Propagation Network" that processes video frames in an adaptive manner. The model is applied online: it propagates information forward without the need to access future frames. In particular we combine two components, a temporal bilateral network for dense and video adaptive filtering, followed by a spatial network to refine features and increased flexibility. We present experiments on video object segmentation and semantic video segmentation and show increased performance comparing to the best previous task-specific methods, while having favorable runtime. Additionally we demonstrate our approach on an example regression task of color propagation in a grayscale video.
**********

Global Hypothesis Generation for 6D Object Pose Estimation
Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 462-471
This paper addresses the task of estimating the 6D-pose of a known 3D object from a single RGB-D image. Most modern approaches solve this task in three steps: i) compute local features; ii) generate a pool of pose-hypotheses; iii) select and refine a pose from the pool. This work focuses on the second step. While all existing approaches generate the hypotheses pool via local reasoning, e.g. RANSAC or Hough-Voting, we are the first to show that global reasoning is beneficial at this stage. In particular, we formulate a novel fully-connected Conditional Random Field (CRF) that outputs a very small number of pose-hypotheses. Despite the potential functions of the CRF being non-Gaussian, we give a new, efficient two-step optimization procedure, with some guarantees for optimality. We utilize our global hypotheses generation procedure to produce results that exceed state-of-the-art for the challenging "Occluded Object Dataset".
**********

Dilated Residual Networks
Fisher Yu, Vladlen Koltun, Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 472-480
Convolutional networks for image classification progressively reduce resolution until the image is represented by tiny feature maps in which the spatial structure of the scene is no longer discernible. Such loss of spatial acuity can limit image classification accuracy and complicate the transfer of the model to downstream applications that require detailed scene understanding. These problems can be alleviated by dilation, which increases the resolution of output feature maps without reducing the receptive field of individual neurons. We show that dilated residual networks (DRNs) outperform their non-dilated counterparts in image classification without increasing the model's depth or complexity. We then study gridding artifacts introduced by dilation, develop an approach to removing these artifacts (`degridding'), and show that this further increases the performance of DRNs. In addition, we show that the accuracy advantage of DRNs is further magnified in downstream applications such as object localization and semantic segmentation.
**********

Robust Interpolation of Correspondences for Large Displacement Optical Flow
Yinlin Hu, Yunsong Li, Rui Song; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 481-489
The interpolation of correspondences (EpicFlow) was widely used for optical flow estimation in most-recent works. It has the advantage of edge-preserving and ef

ficiency. However, it is vulnerable to input matching noise, which is inevitable in modern matching techniques. In this paper, we present a Robust Interpolation method of Correspondences (called RicFlow) to overcome the weakness. First, the scene is over-segmented into superpixels to revitalize an early idea of piecewise flow model. Then, each model is estimated robustly from its support neighbors based on a graph constructed on superpixels. We propose a propagation mechanism among the pieces in the estimation of models. The propagation of models is significantly more efficient than the independent estimation of each model, yet retains the accuracy. Extensive experiments on three public datasets demonstrate that RicFlow is more robust than EpicFlow, and it outperforms state-of-the-art methods.

**********

Supervising Neural Attention Models for Video Captioning by Human Gaze Data

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, Gunhee Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 490-498

The attention mechanisms in deep neural networks are inspired by human's attention that sequentially focuses on the most relevant parts of the information over time to generate prediction output. The attention parameters in those models are implicitly trained in an end-to-end manner, yet there have been few trials to explicitly incorporate human gaze tracking to supervise the attention models. In this paper, we investigate whether attention models can benefit from explicit human gaze labels, especially for the task of video captioning. We collect a new dataset called VAS, consisting of movie clips, and corresponding multiple descriptive sentences along with human gaze tracking data. We propose a video captioning model named Gaze Encoding Attention Network (GEAN) that can leverage gaze tracking information to provide the spatial and temporal attention for sentence generation. Through evaluation of language similarity metrics and human assessment via Amazon mechanical Turk, we demonstrate that spatial attentions guided by human gaze data indeed improve the performance of multiple captioning methods. Moreover, we show that the proposed approach achieves the state-of-the-art performance for both gaze prediction and video captioning not only in our VAS dataset but also in standard datasets (e.g. LSMDC and Hollywood2)

**********

Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks

Hongsong Wang, Liang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 499-508

Recently, skeleton based action recognition gains more popularity due to cost-effective depth sensors coupled with real-time skeleton estimation algorithms. Traditional approaches based on handcrafted features are limited to represent the complexity of motion patterns. Recent methods that use Recurrent Neural Networks (RNN) to handle raw skeletons only focus on the contextual dependency in the temporal domain and neglect the spatial configurations of articulated skeletons. In this paper, we propose a novel two-stream RNN architecture to model both temporal dynamics and spatial configurations for skeleton based action recognition. We explore two different structures for the temporal stream: stacked RNN and hierarchical RNN. Hierarchical RNN is designed according to human body kinematics. We also propose two effective methods to model the spatial structure by converting the spatial graph into a sequence of joints. To improve generalization of our model, we further exploit 3D transformation based data augmentation techniques including rotation and scaling transformation to transform the 3D coordinates of skeletons during training. Experiments on 3D action recognition benchmark datasets show that our method brings a considerable improvement for a variety of actions, i.e., generic actions, interaction activities and gestures.

**********

Self-Learning Scene-Specific Pedestrian Detectors Using a Progressive Latent Model

Qixiang Ye, Tianliang Zhang, Wei Ke, Qiang Qiu, Jie Chen, Guillermo Sapiro, Baochang Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Re

cognition (CVPR), 2017, pp. 509-518
In this paper, a self-learning approach is proposed towards solving scene-specif ic pedestrian detection problem without any human' annotation involved. The self -learning approach is deployed as progressive steps of object discovery, object enforcement, and label propagation. In the learning procedure, object locations in each frame are treated as latent variables that are solved with a progressive latent model (PLM). Compared with conventional latent models, the proposed PLM incorporates a spatial regularization term to reduce ambiguities in object propo sals and to enforce object localization, and also a graph-based label propagatio n to discover harder instances in adjacent frames. With the difference of convex (DC) objective functions, PLM can be efficiently optimized with a concave-conve x programming and thus guaranteeing the stability of self-learning. Extensive ex periments demonstrate that even without annotation the proposed self-learning ap proach outperforms weakly supervised learning approaches, while achieving compar able performance with transfer learning and fully supervised approaches.
**********

Oriented Response Networks
Yanzhao Zhou, Qixiang Ye, Qiang Qiu, Jianbin Jiao; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 519-528
Deep Convolution Neural Networks (DCNNs) are capable of learning unprecedentedly effective image representations. However, their ability in handling significant local and global image rotations remains limited. In this paper, we propose Act ive Rotating Filters (ARFs) that actively rotate during convolution and produce feature maps with location and orientation explicitly encoded. An ARF acts as a virtual filter bank containing the filter itself and its multiple unmaterialised rotated versions. During back-propagation, an ARF is collectively updated using errors from all its rotated versions. DCNNs using ARFs, referred to as Oriented Response Networks (ORNs), can produce within-class rotation-invariant deep feat ures while maintaining inter-class discrimination for classification tasks. The oriented response produced by ORNs can also be used for image and object orienta tion estimation tasks. Over multiple state-of-the-art DCNN architectures, such a s VGG, ResNet, and STN, we consistently observe that replacing regular filters w ith the proposed ARFs leads to significant reduction in the number of network pa rameters and improvement in classification performance. We report the best resul ts on several commonly used benchmarks.
**********

Video Acceleration Magnification
Yichao Zhang, Silvia L. Pintea, Jan C. van Gemert; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 529-537
The ability to amplify or reduce subtle image changes over time is useful in con texts such as video editing, medical video analysis, product quality control and sports. In these contexts there is often large motion present which severely di storts current video amplification methods that magnify change linearly. In this work we propose a method to cope with large motions while still magnifying smal l changes. We make the following two observations: i) large motions are linear o n the temporal scale of the small changes; ii) small changes deviate from this l inearity. We ignore linear motion and propose to magnify acceleration. Our metho d is pure Eulerian and does not require any optical flow, temporal alignment or region annotations. We link temporal second-order derivative filtering to spatia l acceleration magnification. We apply our method to moving objects where we sho w motion magnification and color magnification. We provide quantitative as well as qualitative evidence for our method while comparing to the state-of-the-art.
**********

IRINA: Iris Recognition (Even) in Inaccurately Segmented Data
Hugo Proenca, Joao C. Neves; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2017, pp. 538-547
The effectiveness of current iris recognition systems depends on the accurate se gmentation and parameterisation of the iris boundaries, as failures at this poin t misalign the coefficients of the biometric signatures. This paper describes IR INA, an algorithm for Iris Recognition that is robust against INAccurately segme

nted samples, which makes it a good candidate to work in poor-quality data. The process is based in the concept of "corresponding" patch between pairs of images , that is used to estimate the posterior probabilities that patches regard the s ame biological region, even in case of segmentation errors and non-linear textur e deformations. Such information enables to infer a free-form deformation field (2D registration vectors) between images, whose first and second-order statistic s provide effective biometric discriminating power. Extensive experiments were c arried out in four datasets (CASIA-IrisV3-Lamp, CASIA-IrisV4-Lamp, CASIA-IrisV4-Thousand and WVU) and show that IRINA not only achieves state-of-the-art perform ance in good quality data, but also handles effectively severe segmentation erro rs and large differences in pupillary dilation / constriction.

**********

Forecasting Human Dynamics From Static Images

Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, Jia Deng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 548 -556

This paper presents the first study on forecasting human dynamics from static im ages. The problem is to input a single RGB image and generate a sequence of upco ming human body poses in 3D. To address the problem, we propose the 3D Pose Fore casting Network (3D-PFNet). Our 3D-PFNet integrates recent advances on single-im age human pose estimation and sequence prediction, and converts the 2D predictio ns into 3D space. We train our 3D-PFNet using a three-step training strategy to leverage a diverse source of training data, including image and video based huma n pose datasets and 3D motion capture (MoCap) data. We demonstrate competitive p erformance of our 3D-PFNet on 2D pose forecasting and 3D structure recovery thro ugh quantitative and qualitative results.

**********

Discriminative Bimodal Networks for Visual Localization and Detection With Natur al Language Queries

Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, Honglak Lee; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 557-566

Associating image regions with text queries has been recently explored as a new way to bridge visual and linguistic representations. A few pioneering approaches have been proposed based on recurrent neural language models trained generative ly (e.g., generating captions), but achieving somewhat limited localization accu racy. To better address natural-language-based visual entity localization, we pr opose a discriminative approach. We formulate a discriminative bimodal neural ne twork (DBNet), which can be trained by a classifier with extensive use of negati ve samples. Our training objective encourages better localization on single imag es, incorporates text phrases in a broad range, and properly pairs image regions with text phrases into positive and negative examples. Experiments on the Visua l Genome dataset demonstrate the proposed DBNet significantly outperforms previo us state-of-the-art methods both for localization on single images and for detec tion on multiple images. We we also establish an evaluation protocol for natural -language visual detection.

**********

A Linear Extrinsic Calibration of Kaleidoscopic Imaging System From Single 3D Po int

Kosuke Takahashi, Akihiro Miyata, Shohei Nobuhara, Takashi Matsuyama; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 , pp. 567-575

This paper proposes a new extrinsic calibration of kaleidoscopic imaging system by estimating normals and distances of the mirrors. The problem to be solved in this paper is a simultaneous estimation of all mirror parameters consistent thro ughout multiple reflections. Unlike conventional methods utilizing a pair of dir ect and mirrored images of a reference 3D object to estimate the parameters on a per-mirror basis, our method renders the simultaneous estimation problem into s olving a linear set of equations. The key contribution of this paper is to intro duce a linear estimation of multiple mirror parameters from kaleidoscopic 2D pro

jections of a single 3D point of unknown geometry. Evaluations with synthesized and real images demonstrate the performance of the proposed algorithm in compari son with conventional methods.
**********

Efficient Multiple Instance Metric Learning Using Weakly Supervised Data
Marc T. Law, Yaoliang Yu, Raquel Urtasun, Richard S. Zemel, Eric P. Xing; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 576-584
We consider learning a distance metric in a weakly supervised setting where "bag s" (or sets) of instances are labeled with "bags" of labels. A general approach is to formulate the problem as a Multiple Instance Learning (MIL) problem where the metric is learned so that the distances between instances inferred to be sim ilar are smaller than the distances between instances inferred to be dissimilar. Classic approaches alternate the optimization over the learned metric and the a ssignment of similar instances. In this paper, we propose an efficient method th at jointly learns the metric and the assignment of instances. In particular, our model is learned by solving an extension of k-means for MIL problems where inst ances are assigned to categories depending on annotations provided at bag-level. Our learning algorithm is much faster than existing metric learning methods for MIL problems and obtains state-of-the-art recognition performance in automated image annotation and instance classification for face identification.
**********

Asynchronous Temporal Fields for Action Recognition
Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, Abhinav Gupta; Proceedings o f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, p p. 585-594
Actions are more than just movements and trajectories: we cook to eat and we hol d a cup to drink from it. A thorough understanding of videos requires going beyo nd appearance modeling and necessitates reasoning about the sequence of activiti es, as well as the higher-level constructs such as intentions. But how do we mod el and reason about these? We propose a fully-connected temporal CRF model for r easoning over various aspects of activities that includes objects, actions, and intentions, where the potentials are predicted by a deep network. End-to-end tra ining of such structured models is a challenging endeavor: For inference and lea rning we need to construct mini-batches consisting of whole videos, leading to m ini-batches with only a few videos. This causes high-correlation between data po ints leading to breakdown of the backprop algorithm. To address this challenge, we present an asynchronous variational inference method that allows efficient en d-to-end training. Our method achieves a classification mAP of 22.4% on the Char ades benchmark, outperforming the state-of-the-art (17.2% mAP), and offers equal gains on the task of temporal localization.
**********

Scene Flow to Action Map: A New Representation for RGB-D Based Action Recognitio n With Convolutional Neural Networks
Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, Philip Ogunbona; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 595-604
Scene flow describes the motion of 3D objects in real world and potentially coul d be the basis of a good feature for 3D action recognition. However, its use for action recognition, especially in the context of convolutional neural networks (ConvNets), has not been previously studied. In this paper, we propose the extra ction and use of scene flow for action recognition from RGB-D data. Previous wor ks have considered the depth and RGB modalities as separate channels and extract features for later fusion. We take a different approach and consider the modali ties as one entity, thus allowing feature extraction for action recognition at t he beginning. Two key questions about the use of scene flow for action recogniti on are addressed: how to organize the scene flow vectors and how to represent th e long term dynamics of videos based on scene flow. In order to calculate the sc ene flow correctly on the available datasets, we propose an effective self-calib ration method to align the RGB and depth data spatially without knowledge of the

camera parameters. Based on the scene flow vectors, we propose a new representation, namely, Scene Flow to Action Map (SFAM), that describes several long term spatio-temporal dynamics for action recognition. We adopt a channel transform kernel to transform the scene flow vectors to an optimal color space analogous to RGB. This transformation takes better advantage of the trained ConvNets models over ImageNet. Experimental results indicate that this new representation can surpass the performance of state-of-the-art methods on two large public datasets.
**********

A Point Set Generation Network for 3D Object Reconstruction From a Single Image
Haoqiang Fan, Hao Su, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 605-613
Generation of 3D data by deep neural network has been attracting increasing attention in the research community. The majority of extant works resort to regular representations such as volumetric grids or collection of images; however, these representations obscure the natural invariance of 3D shapes under geometric transformations, and also suffer from a number of other issues. In this paper we address the problem of 3D reconstruction from a single image, generating a straight-forward form of output -- point cloud coordinates. Along with this problem arises a unique and interesting issue, that the groundtruth shape for an input image may be ambiguous. Driven by this unorthordox output form and the inherent ambiguity in groundtruth, we design architecture, loss function and learning paradigm that are novel and effective. Our final solution is a conditional shape sampler, capable of predicting multiple plausible 3D point clouds from an input image. In experiments not only can our system outperform state-of-the-art methods on single image based 3D reconstruction benchmarks; but it also shows strong performance for 3D shape completion and promising ability in making multiple plausible predictions.
**********

Automatic Discovery, Association Estimation and Learning of Semantic Attributes for a Thousand Categories
Ziad Al-Halah, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 614-623
Attribute-based recognition models, due to their impressive performance and their ability to generalize well on novel categories, have been widely adopted for many computer vision applications. However, usually both the attribute vocabulary and the class-attribute associations have to be provided manually by domain experts or large number of annotators. This is very costly and not necessarily optimal regarding recognition performance, and most importantly, it limits the applicability of attribute-based models to large scale data sets. To tackle this problem, we propose an end-to-end unsupervised attribute learning approach. We utilize online text corpora to automatically discover a salient and discriminative vocabulary that correlates well with the human concept of semantic attributes. Moreover, we propose a deep convolutional model to optimize class-attribute associations with a linguistic prior that accounts for noise and missing data in text. In a thorough evaluation on ImageNet, we demonstrate that our model is able to efficiently discover and learn semantic attributes at a large scale. Furthermore, we demonstrate that our model outperforms the state-of-the-art in zero-shot learning on three data sets: ImageNet, Animals with Attributes and aPascal/aYahoo. Finally, we enable attribute-based learning on ImageNet and will share the attributes and associations for future research.
**********

Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution
Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 624-632
Convolutional neural networks have recently demonstrated high-quality reconstruction for single-image super-resolution. In this paper, we propose the Laplacian Pyramid Super-Resolution Network (LapSRN) to progressively reconstruct the sub-band residuals of high-resolution images. At each pyramid level, our model takes coarse-resolution feature maps as input, predicts the high-frequency residuals,

and uses transposed convolutions for upsampling to the finer level. Our method d
oes not require the bicubic interpolation as the pre-processing step and thus dr
amatically reduces the computational complexity. We train the proposed LapSRN wi
th deep supervision using a robust Charbonnier loss function and achieve high-qu
ality reconstruction. Furthermore, our network generates multi-scale predictions
 in one feed-forward pass through the progressive reconstruction, thereby facili
tates resource-aware applications. Extensive quantitative and qualitative evalua
tions on benchmark datasets show that the proposed algorithm performs favorably
against the state-of-the-art methods in terms of speed and accuracy.
**********

Scene Parsing Through ADE20K Dataset
Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, Antonio Torral
ba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2017, pp. 633-641
Scene parsing, or recognizing and segmenting objects and stuff in an image, is o
ne of the key problems in computer vision. Despite the community's efforts in da
ta collection, there are still few image datasets covering a wide range of scene
s and object categories with dense and detailed annotations for scene parsing. I
n this paper, we introduce and analyze the ADE20K dataset, spanning diverse anno
tations of scenes, objects, parts of objects, and in some cases even parts of pa
rts. A scene parsing benchmark is built upon the ADE20K with 150 object and stuf
f classes included. Several segmentation baseline models are evaluated on the be
nchmark. A novel network design called Cascade Segmentation Module is proposed t
o parse a scene into stuff, objects, and object parts in a cascade and improve o
ver the baselines. We further show that the trained scene parsing networks can l
ead to applications such as image content removal and scene synthesis(Dataset an
d pretrained models are available at http://groups.csail.mit.edu/vision/datasets
/ADE20K/).
**********

WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, P
ointwise Localization and Segmentation
Thibaut Durand, Taylor Mordan, Nicolas Thome, Matthieu Cord; Proceedings of the
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 642
-651
This paper introduces WILDCAT, a deep learning method which jointly aims at alig
ning image regions for gaining spatial invariance and learning strongly localize
d features. Our model is trained using only global image labels and is devoted t
o three main visual recognition tasks: image classification, weakly supervised o
bject localization and semantic segmentation. WILDCAT extends state-of-the-art C
onvolutional Neural Networks at three main levels: the use of Fully Convolutiona
l Networks for maintaining spatial resolution, the explicit design in the networ
k of local features related to different class modalities, and a new way to pool
 these features to provide a global image prediction required for weakly supervi
sed training. Extensive experiments show that our model significantly outperform
s state-of-the-art methods.
**********

PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation
Charles R. Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652-660
Point cloud is an important type of geometric data structure. Due to its irregul
ar format, most researchers transform such data to regular 3D voxel grids or col
lections of images. This, however, renders data unnecessarily voluminous and cau
ses issues. In this paper, we design a novel type of neural network that directl
y consumes point clouds, which well respects the permutation invariance of point
s in the input. Our network, named PointNet, provides a unified architecture for
 applications ranging from object classification, part segmentation, to scene se
mantic parsing. Though simple, PointNet is highly efficient and effective. Empir
ically, it shows strong performance on par or even better than state of the art.
 Theoretically, we provide analysis towards understanding of what the network ha
s learnt and why the network is robust with respect to input perturbation and co

rruption.
**********
L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space
Yurun Tian, Bin Fan, Fuchao Wu; Proceedings of the IEEE Conference on Computer V ision and Pattern Recognition (CVPR), 2017, pp. 661-669
The research focus of designing local patch descriptors has gradually shifted fr om handcrafted ones (e.g., SIFT) to learned ones. In this paper, we propose to l earn high per- formance descriptor in Euclidean space via the Convolu- tional Ne ural Network (CNN). Our method is distinctive in four aspects: (i) We propose a progressive sampling strat- egy which enables the network to access billions of train- ing samples in a few epochs. (ii) Derived from the ba- sic concept of loc al patch matching problem, we empha- size the relative distance between descript ors. (iii) Extra supervision is imposed on the intermediate feature maps. (iv) C ompactness of the descriptor is taken into account. The proposed network is name d as L2-Net since the out- put descriptor can be matched in Euclidean space by L 2 distance. L2-Net achieves state-of-the-art performance on the Brown datasets [ 16], Oxford dataset [18] and the new- ly proposed Hpatches dataset [11]. The goo d generaliza- tion ability shown by experiments indicates that L2-Net can serve as a direct substitution of the existing handcrafted de- scriptors. The pre-trai ned L2-Net is publicly available.
**********
Video Frame Interpolation via Adaptive Convolution
Simon Niklaus, Long Mai, Feng Liu; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 670-679
Video frame interpolation typically involves two steps: motion estimation and pi xel synthesis. Such a two-step approach heavily depends on the quality of motion estimation. This paper presents a robust video frame interpolation method that combines these two steps into a single process. Specifically, our method conside rs pixel synthesis for the interpolated frame as local convolution over two inpu t frames. The convolution kernel captures both the local motion between the inpu t frames and the coefficients for pixel synthesis. Our method employs a deep ful ly convolutional neural network to estimate a spatially-adaptive convolution ker nel for each pixel. This deep neural network can be directly trained end to end using widely available video data without any difficult-to-obtain ground-truth d ata like optical flow. Our experiments show that the formulation of video interp olation as a single convolution process allows our method to gracefully handle c hallenges like occlusion, blur, and abrupt brightness change and enables high-qu ality video frame interpolation.
**********
Crossing Nets: Combining GANs and VAEs With a Shared Latent Space for Hand Pose Estimation
Chengde Wan, Thomas Probst, Luc Van Gool, Angela Yao; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 680-689
State-of-the-art methods for 3D hand pose estimation from depth images require l arge amounts of annotated training data. We propose modelling the statistical re lationship of 3D hand poses and corresponding depth images using two deep genera tive models with a shared latent space. By design, our architecture allows for l earning from unlabeled image data in a semi-supervised manner. Assuming a one-to -one mapping between a pose and a depth map, any given point in the shared laten t space can be projected into both a hand pose or into a corresponding depth map . Regressing the hand pose can then be done by learning a discriminator to estim ate the posterior of the latent pose given some depth map. To prevent over-fitti ng and to better exploit unlabeled depth maps, the generator and discriminator a re trained jointly. At each iteration, the generator is updated with the back-pr opagated gradient from the discriminator to synthesize realistic depth maps of t he articulated hand, while the discriminator benefits from an augmented training set of synthesized samples and unlabeled depth maps. The proposed discriminator network architecture is highly efficient and runs at 90fps on the CPU with accu racies comparable or better than state-of-art on 3 publicly available benchmarks .

**********
Attention-Aware Face Hallucination via Deep Reinforcement Learning
Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, Guanbin Li; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6 90-698
Face hallucination is a domain-specific super-resolution problem with the goal t o generate high-resolution (HR) faces from low-resolution (LR) input images. In contrast to existing methods that often learn a single patch-to-patch mapping fr om LR to HR images and are regardless of the contextual interdependency between patches, we propose a novel Attention-aware Face Hallucination (Attention-FH) fr amework which resorts to deep reinforcement learning for sequentially discoverin g attended patches and then performing the facial part enhancement by fully expl oiting the global interdependency of the image. Specifically, in each time step, the recurrent policy network is proposed to dynamically specify a new attended region by incorporating what happened in the past. The state (i.e., face halluci nation result for the whole image) can thus be exploited and updated by the loca l enhancement network on the selected region. The Attention-FH approach jointly learns the recurrent policy network and local enhancement network through maximi zing the long-term reward that reflects the hallucination performance over the w hole image. Therefore, our proposed Attention-FH is capable of adaptively person alizing an optimal searching path for each face image according to its own chara cteristic. Extensive experiments show our approach significantly surpasses the s tate-of-the-arts on in-the-wild faces with large pose and illumination variation s.The state (i.e., face hallucination result for the whole image) can thus be ex ploited and updated by the local enhancement network on the selected region. The Attention-FH approach jointly learns the recurrent policy network and local enh ancement network through maximizing the long-term reward that reflects the hallu cination performance over the whole image. Therefore, our proposed Attention-FH is capable of adaptively personalizing an optimal searching path for each face i mage according to its own characteristic. Extensive experiments show our approac h significantly surpasses the state-of-the-arts on in-the-wild faces with large pose and illumination variations.
**********
Neural Scene De-Rendering
Jiajun Wu, Joshua B. Tenenbaum, Pushmeet Kohli; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 699-707
e study the problem of holistic scene understanding. We would like to obtain a c ompact, expressive, and interpretable representation of scenes that encodes info rmation such as the number of objects and their categories, poses, positions, et c. Such a representation would allow us to reason about and even reconstruct or manipulate elements of the scene. Previous works have used encoder-decoder based neural architectures to learn image representations; however, representations o btained in this way are typically uninterpretable, or only explain a single obje ct in the scene.  In this work, we propose a new approach to learn an interpreta ble distributed representation of scenes. Our approach employs a deterministic r endering function as the decoder, mapping a naturally structured and disentangle d scene description, which we named scene XML, to an image. By doing so, the enc oder is forced to perform the inverse of the rendering operation (a.k.a. de-rend ering) to transform an input image to the structured scene XML that the decoder used to produce the image. We use a object proposal based encoder that is traine d by minimizing both the supervised prediction and the unsupervised reconstructi on errors. Experiments demonstrate that our approach works well on scene de-rend ering with two different graphics engines, and our learned representation can be easily adapted for a wide range of applications like image editing, inpainting, visual analogy-making, and image captioning.
**********
Deep TEN: Texture Encoding Network
Hang Zhang, Jia Xue, Kristin Dana; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 708-717
We propose a Deep Texture Encoding Network (TEN) with a novel Encoding Layer int

egrated on top of convolutional layers, which ports the entire dictionary learni ng and encoding pipeline into a single model. Current methods build from distinc t components, using standard encoders with separate off-the-shelf features such as such as SIFT descriptors or pre-trained CNN features for material recognition . Our new approach provides an end-to-end learning framework, where the inherent visual vocabularies are learned directly from the loss function. That is, the f eatures, dictionaries and the encoding representation for the classifier are all learned simultaneously. The representation is orderless and therefore is partic ularly useful for material and texture recognition. This Encoding Layer generali zes robust residual encoders such as VLAD and Fisher Vectors, and has the proper ty of discarding domain specific information which makes the learned convolution al features easier to transfer. Additionally, joint training using multiple data sets of varied sizes and class labels is supported resulting in increased recogn ition performance. The experimental results show superior performance as compare d to state-of-the-art methods using gold-standard databases such as MINC-2500, F licker Material Database, KTH-TIPS-2b, and a new ground terrain multiview databa se. The source code for the complete system will be publicly available upon publ ication.

**********

PolyNet: A Pursuit of Structural Diversity in Very Deep Networks

Xingcheng Zhang, Zhizhong Li, Chen Change Loy, Dahua Lin; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 718-72 6

A number of studies have shown that increasing the depth or width of convolution al networks is a rewarding approach to improve the performance of image recognit ion. In our study, however, we observed difficulties along both directions. On o ne hand, the pursuit for very deep networks is met with a diminishing return and increased training difficulty; on the other hand, widening a network would resu lt in a quadratic growth in both computational cost and memory demand. These dif ficulties motivate us to explore structural diversity in designing deep networks , a new dimension beyond just depth and width. Specifically, we present a new fa mily of modules, namely the PolyInception, which can be flexibly inserted in iso lation or in a composition as replacements of different parts of a network. Choo sing PolyInception modules with the guidance of architectural efficiency can imp rove the expressive power while preserving comparable computational cost. The Ve ry Deep PolyNet, designed following this direction, demonstrates substantial imp rovements over the state-of-the-art on the ILSVRC 2012 benchmark. Compared to In ception-ResNet-v2, it reduces the top-5 validation error on single crops from 4. 9% to 4.25%, and that on multi-crops from 3.7% to 3.45%.

**********

Object Detection in Videos With Tubelet Proposal Networks

Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni tion (CVPR), 2017, pp. 727-735

Object detection in videos has drawn increasing attention recently with the intr oduction of the large-scale ImageNet VID dataset. Different from object detectio n in static images, temporal information in videos is vital for object detection . To fully utilize temporal information, state-of-the-art methods are based on s patiotemporal tubelets, which are essentially sequences of associated bounding b oxes across time. However, the existing methods have major limitations in genera ting tubelets in terms of quality and efficiency. Motion-based methods are able to obtain dense tubelets efficiently, but the lengths are generally only several frames, which is not optimal for incorporating long-term temporal information. Appearance-based methods, usually involving generic object tracking, could gener ate long tubelets, but are usually computationally expensive. In this work, we p ropose a framework for object detection in videos, which consists of a novel tub elet proposal network to efficiently generate spatiotemporal proposals, and a Lo ng Short-term Memory (LSTM) network that incorporates temporal information from tubelet proposals for achieving high object detection accuracy in videos. Experi ments on the large-scale ImageNet VID dataset demonstrate the effectiveness of t

he proposed framework for object detection in videos.
**********

AMVH: Asymmetric Multi-Valued Hashing
Cheng Da, Shibiao Xu, Kun Ding, Gaofeng Meng, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 736-744
Most existing hashing methods resort to binary codes for similarity search, owing to the high efficiency of computation and storage. However, binary codes lack enough capability in similarity preservation, resulting in less desirable performance. To address this issue, we propose an asymmetric multi-valued hashing method supported by two different non-binary embeddings. (1) A real-valued embedding is used for representing the newly-coming query. (2) A multi-integer-embedding is employed for compressing the whole database, which is modeled by binary sparse representation with fixed sparsity. With these two non-binary embeddings, the similarities between data points can be preserved precisely. To perform meaningful asymmetric similarity computation for efficient semantic search, these embeddings are jointly learnt by preserving the label-based similarity. Technically, this results in a mixed integer programming problem, which is efficiently solved by alternative optimization. Extensive experiments on three multilabel datasets demonstrate that our approach not only outperforms the existing binary hashing methods in search accuracy, but also retains their query and storage efficiency.
**********

Real-Time 3D Model Tracking in Color and Depth on a Single CPU Core
Wadim Kehl, Federico Tombari, Slobodan Ilic, Nassir Navab; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 745-753
We present a novel method to track 3D models in color and depth data. To this end, we introduce approximations that accelerate the state-of-the-art in region-based tracking by an order of magnitude while retaining similar accuracy. Furthermore, we show how the method can be made more robust in the presence of depth data and consequently formulate a new joint contour and ICP tracking energy. We present better results than the state-of-the-art while being much faster then most other methods and achieving all of the above on a single CPU core.
**********

Weakly Supervised Action Learning With RNN Based Fine-To-Coarse Modeling
Alexander Richard, Hilde Kuehne, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 754-763
We present an approach for weakly supervised learning of human actions. Given a set of videos and an ordered list of the occurring actions, the goal is to infer start and end frames of the related action classes within the video and to train the respective action classifiers without any need for hand labeled frame boundaries. To address this task, we propose a combination of a discriminative representation of subactions, modeled by a recurrent neural network, and a coarse probabilistic model to allow for a temporal alignment and inference over long sequences. While this system alone already generates good results, we show that the performance can be further improved by approximating the number of subactions to the characteristics of the different action classes. To this end, we adapt the number of subaction classes by iterating realignment and reestimation during training. The proposed system is evaluated on two benchmark datasets, the Breakfast and the Hollywood extended dataset, showing a competitive performance on various weak learning tasks such as temporal action segmentation and action alignment.
**********

Differential Angular Imaging for Material Recognition
Jia Xue, Hang Zhang, Kristin Dana, Ko Nishino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 764-773
Material recognition for real-world outdoor surfaces has become increasingly important for computer vision to support its operation "in the wild." Computational surface modeling that underlies material recognition has transitioned from reflectance modeling using in-lab controlled radiometric measurements to image-based representations based on internet-mined images of materials captured in the sce

ne. We propose to take a middle-ground approach for material recognition that ta kes advantage of both rich radiometric cues and flexible image capture. We reali ze this by developing a framework for differential angular imaging, where small angular variations in image capture provide an enhanced appearance representatio n and significant recognition improvement. We build a large-scale material datab ase, Ground Terrain in Outdoor Scenes (GTOS) database, geared towards real use f or autonomous agents. The database consists of over 30,000 images covering 40 cl asses of outdoor ground terrain under varying weather and lighting conditions. W e develop a novel approach for material recognition called a Differential An- gu lar Imaging Network (DAIN) to fully leverage this large dataset. With this novel network architecture, we extract characteristics of materials encoded in the an gular and spatial gradients of their appearance. Our results show that DAIN achi eves recognition performance that surpasses single view or coarsely quantized mu ltiview images. These results demonstrate the effectiveness of differential angu lar imaging as a means for flexible, in-place material recognition.

**********

Forecasting Interactive Dynamics of Pedestrians With Fictitious Play

Wei-Chiu Ma, De-An Huang, Namhoon Lee, Kris M. Kitani; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 774-782

We develop predictive models of pedestrian dynamics by encoding the coupled natu re of multi-pedestrian interaction using game theory and deep learning-based vis ual analysis to estimate person-specific behavior parameters. We focus on predic tive models since they are important for developing interactive autonomous syste ms (e.g., autonomous cars, home robots, smart homes) that can understand differe nt human behavior and pre-emptively respond to future human actions. Building pr edictive models for multi-pedestrian interactions however, is very challenging d ue to two reasons: (1) the dynamics of interaction are complex interdependent pr ocesses, where the decision of one person can affect others; and (2) dynamics ar e variable, where each person may behave differently (e.g., an older person may walk slowly while the younger person may walk faster). To address these challeng es, we utilize concepts from game theory to model the intertwined decision makin g process of multiple pedestrians and use visual classifiers to learn a mapping from pedestrian appearance to behavior parameters. We evaluate our proposed mode l on several public multiple pedestrian interaction video datasets. Results show that our strategic planning model predicts and explains human interactions 25% better when compared to a state-of-the-art activity forecasting method.

**********

Real-Time Neural Style Transfer for Videos

Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, Wei Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition (CVPR), 2017, pp. 783-791

Recent research endeavors have shown the potential of using feed-forward convolu tional neural networks to accomplish fast style transfer for images. In this wor k, we take one step further to explore the possibility of exploiting a feed-forw ard network to perform style transfer for videos and simultaneously maintain tem poral consistency among stylized video frames. Our feed-forward network is train ed by enforcing the outputs of consecutive frames to be both well stylized and t emporally consistent. More specifically, a hybrid loss is proposed to capitalize on the content information of input frames, the style information of a given st yle image, and the temporal information of consecutive frames. To calculate the temporal loss during the training stage, a novel two-frame synergic training mec hanism is proposed. Compared with directly applying an existing image style tran sfer method to videos, our proposed method employs the trained network to yield temporally consistent stylized videos which are much more visually pleasant. In contrast to the prior video style transfer method which relies on time-consuming optimization on the fly, our method runs in real time while generating competit ive visual results.

**********

Incremental Kernel Null Space Discriminant Analysis for Novelty Detection

Juncheng Liu, Zhouhui Lian, Yi Wang, Jianguo Xiao; Proceedings of the IEEE Confe

rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 792-800
Novelty detection, which aims to determine whether a given data belongs to any c ategory of training data or not, is considered to be an important and challengin g problem in areas of Pattern Recognition, Machine Learning, etc. Recently, kern el null space method (KNDA) was reported to have state-of-the-art performance in novelty detection. However, KNDA is hard to scale up because of its high comput ational cost. With the ever-increasing size of data, accelerating the implementi ng speed of KNDA is desired and critical. Moreover, it becomes incapable when th ere exist successively injected data. To address these issues, we propose the In cremental Kernel Null Space based Discriminant Analysis (IKNDA) algorithm. The k ey idea is to extract new information brought by newly-added samples and integra te it with the existing model by an efficient updating scheme. Experiments condu cted on two publicly-available datasets demonstrate that the proposed IKNDA yiel ds comparable performance as the batch KNDA yet significantly reduces the comput ational complexity, and our IKNDA based novelty detection methods markedly outpe rform approaches using deep neural network (DNN) classifiers. This validates the superiority of our IKNDA against the state of the art in novelty detection for large-scale data.
**********
Self-Calibration-Based Approach to Critical Motion Sequences of Rolling-Shutter Structure From Motion
Eisuke Ito, Takayuki Okatani; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2017, pp. 801-809
In this paper we consider critical motion sequences (CMSs) of rolling-shutter (R S) SfM. Employing an RS camera model with linearized pure rotation, we show that the RS distortion can be approximately expressed by two internal parameters of an "imaginary" camera plus one-parameter nonlinear transformation similar to len s distortion. We then reformulate the problem as self-calibration of the imagina ry camera, in which its skew and aspect ratio are unknown and varying in the ima ge sequence. In the formulation, we derive a general representation of CMSs. We also show that our method can explain the CMS that was recently reported in the literature, and then present a new remedy to deal with the degeneracy. Our theor etical results agree well with experimental results; it explains degeneracies ob served when we employ naive bundle adjustment, and how they are resolved by our method.
**********
Recurrent 3D Pose Sequence Machines
Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, Hui Cheng; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 810-81 9
3D human articulated pose recovery from monocular image sequences is very challe nging due to the diverse appearances, viewpoints, occlusions, and also the human 3D pose is inherently ambiguous from the monocular imagery. It is thus critical to exploit rich spatial and temporal long-range dependencies among body joints for accurate 3D pose sequence prediction. Existing approaches usually manually d esign some elaborate prior terms and human body kinematic constraints for captur ing structures, which are often insufficient to exploit all intrinsic structures and not scalable for all scenarios. In contrast, this paper presents a Recurren t 3D Pose Sequence Machine(RPSM) to automatically learn the image-dependent stru ctural constraint and sequence-dependent temporal context by using a multi-stage sequential refinement. At each stage, our RPSM is composed of three modules to predict the 3D pose sequences based on the previously learned 2D pose representa tions and 3D poses: (i) a 2D pose module extracting the image-dependent pose rep resentations, (ii) a 3D pose recurrent module regressing 3D poses and (iii) a fe ature adaption module serving as a bridge between module (i) and (ii) to enable the representation transformation from 2D to 3D domain. These three modules are then assembled into a sequential prediction framework to refine the predicted po ses with multiple recurrent stages. Extensive evaluations on the Human3.6M datas et and HumanEva-I dataset show that our RPSM outperforms all state-of-the-art ap proaches for 3D pose estimation.

**********
Efficient Solvers for Minimal Problems by Syzygy-Based Reduction
Viktor Larsson, Kalle Astrom, Magnus Oskarsson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 820-829
In this paper we study the problem of automatically generating polynomial solvers for minimal problems. The main contribution is a new method for finding small elimination templates by making use of the syzygies (i.e. the polynomial relations) that exist between the original equations. Using these syzygies we can essentially parameterize the set of possible elimination templates. We evaluate our method on a wide variety of problems from geometric computer vision and show improvement compared to both handcrafted and automatically generated solvers. Furthermore we apply our method on two previously unsolved relative orientation problems.
**********
Conditional Similarity Networks
Andreas Veit, Serge Belongie, Theofanis Karaletsos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 830-838
What makes images similar? To measure the similarity between images, they are typically embedded in a feature-vector space, in which their distance preserve the relative dissimilarity. However, when learning such similarity embeddings the simplifying assumption is commonly made that images are only compared to one unique measure of similarity. A main reason for this is that contradicting notions of similarities cannot be captured in a single space. To address this shortcoming, we propose Conditional Similarity Networks (CSNs) that learn embeddings differentiated into semantically distinct subspaces that capture the different notions of similarities. CSNs jointly learn a disentangled embedding where features for different similarities are encoded in separate dimensions as well as masks that select and reweight relevant dimensions to induce a subspace that encodes a specific similarity notion. We show that our approach learns interpretable image representations with visually relevant semantic subspaces. Further, when evaluating on triplet questions from multiple similarity notions our model even outperforms the accuracy obtained by training individual specialized networks for each notion separately.
**********
Learning From Noisy Large-Scale Datasets With Minimal Supervision
Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 839-847
We present an approach to effectively use millions of images with noisy annotations in conjunction with a small subset of cleanly-annotated images to learn powerful image representations. One common approach to combine clean and noisy data is to first pre-train a network using the large noisy dataset and then fine-tune with the clean dataset. We show this approach does not fully leverage the information contained in the clean set. Thus, we demonstrate how to use the clean annotations to reduce the noise in the large dataset before fine-tuning the network using both the clean set and the full set with reduced noise. The approach comprises a multi-task network that jointly learns to clean noisy annotations and to accurately classify images. We evaluate our approach on the recently released Open Images dataset, containing 9 million images, multiple annotations per image and over 6000 unique classes. For the small clean set of annotations we use a quarter of the validation set with 40k images. Our results demonstrate that the proposed approach clearly outperforms direct fine-tuning across all major categories of classes in the Open Image dataset. Further, our approach is particularly effective for a large number of classes with wide range of noise in annotations (20-80% false positive annotations).
**********
Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection
Xiaodan Liang, Lisa Lee, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 848-857

Despite progress in visual perception tasks such as image classification and det
ection, computers still struggle to understand the interdependency of objects in
 the scene as a whole, e.g., relations between objects or their attributes. Exis
ting methods often ignore global context cues capturing the interactions among d
ifferent object instances, and can only recognize a handful of types by exhausti
vely training individual detectors for all possible relationships. To capture su
ch global interdependency, we propose a deep Variation-structured Reinforcement
 Learning (VRL) framework to sequentially discover object relationships and attri
butes in the whole image. First, a directed semantic action graph is built using
 language priors to provide a rich and compact representation of semantic correl
ations between object categories, predicates, and attributes. Next, we use a var
iation-structured traversal over the action graph to construct a small, adaptive
 action set for each step based on the current state and historical actions. In
particular, an ambiguity-aware object mining scheme is used to resolve semantic
ambiguity among object categories that the object detector fails to distinguish.
 We then make sequential predictions using a deep RL framework, incorporating gl
obal context cues and semantic embeddings of previously extracted phrases in the
 state vector. Our experiments on the Visual Relationship Detection (VRD) datase
t and the large-scale Visual Genome dataset validate the superiority of VRL, whi
ch can achieve significantly better detection results on datasets involving thou
sands of relationship and attribute types. We also demonstrate that VRL is able
to predict unseen types embedded in our action graph by learning correlations on
 shared graph nodes.
**********

Convolutional Random Walk Networks for Semantic Image Segmentation
Gedas Bertasius, Lorenzo Torresani, Stella X. Yu, Jianbo Shi; Proceedings of the
 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 85
8-866
Most current semantic segmentation methods rely on fully convolutional networks
(FCNs). However, their use of large receptive fields and many pooling layers cau
se low spatial resolution inside the deep layers. This leads to predictions with
 poor localization around the boundaries. Prior work has attempted to address th
is issue by post-processing predictions with CRFs or MRFs. But such models often
 fail to capture semantic relationships between objects, which causes spatially
disjoint predictions. To overcome these problems, recent methods integrated CRFs
 or MRFs into an FCN framework. The downside of these new models is that they ha
ve much higher complexity than traditional FCNs, which renders training and test
ing more challenging. In this work we introduce a simple, yet effective Convolut
ional Random Walk Network (RWN) that addresses the issues of poor boundary local
ization and spatially fragmented predictions with very little increase in model
complexity. Our proposed RWN jointly optimizes the objectives of pixelwise affin
ity and semantic segmentation. It combines these two objectives via a novel rand
om walk layer that enforces consistent spatial grouping in the deep layers of th
e network. Our RWN is implemented using standard convolution and matrix multipli
cation. This allows an easy integration into existing FCN frameworks and it enab
les end-to-end training of the whole network via standard back-propagation. Our
implementation of RWN requires just 131 additional parameters compared to the tr
aditional FCNs, and yet it consistently produces an improvement over the FCNs on
 semantic segmentation and scene labeling.
**********

Predicting Ground-Level Scene Layout From Aerial Imagery
Menghua Zhai, Zachary Bessinger, Scott Workman, Nathan Jacobs; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 8
67-875
We introduce a novel strategy for learning to extract semantically meaningful fe
atures from aerial imagery. Instead of manually labeling the aerial imagery, we
propose to predict (noisy) semantic features automatically extracted from co-loc
ated ground imagery. Our network architecture takes an aerial image as input, ex
tracts features using a convolutional neural network, and then applies an adapti
ve transformation to map these features into the ground-level perspective. We us

e an end-to-end learning approach to minimize the difference between the semanti c segmentation extracted directly from the ground image and the semantic segment ation predicted solely based on the aerial image.  We show that a model learned using this strategy, with no additional training, is already capable of rough se mantic labeling of aerial imagery. Furthermore, we demonstrate that by finetunin g this model we can achieve more accurate semantic segmentation than two baselin e initialization strategies. We use our network to address the task of estimatin g the geolocation and geo-orientation of a ground image. Finally, we show how fe atures extracted from an aerial image can be used to hallucinate a plausible gro und-level panorama.
**********

Simple Does It: Weakly Supervised Instance and Semantic Segmentation
Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, Bernt Schiele; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 876-885
Semantic labelling and instance segmentation are two tasks that require particul arly costly annotations. Starting from weak supervision in the form of bounding box detection annotations, we propose a new approach that does not require modif ication of the segmentation training procedure. We show that when carefully desi gning the input labels from given bounding boxes, even a single round of trainin g is enough to improve over previously reported weakly supervised results. Overa ll, our weak supervision approach reaches  95% of the quality of the fully super vised model, both for semantic labelling and instance segmentation.
**********

Fast Fourier Color Constancy
Jonathan T. Barron, Yun-Ta Tsai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 886-894
We present Fast Fourier Color Constancy (FFCC), a color constancy algorithm whic h solves illuminant estimation by reducing it to a spatial localization task on a torus. By operating in the frequency domain, FFCC produces lower error rates t han the previous state-of-the-art by 13-20% while being 250-3000 times faster. T his unconventional approach introduces challenges regarding aliasing, directiona l statistics, and preconditioning, which we address. By producing a complete pos terior distribution over illuminants instead of a single illuminant estimate, FF CC enables better training techniques, an effective temporal smoothing technique , and richer methods for error analysis. Our implementation of FFCC runs at  700  frames per second on a mobile device, allowing it to be used as an accurate, re al-time, temporally-coherent automatic white balance algorithm.
**********

Attend to You: Personalized Image Captioning With Context Sequence Memory Networ ks
Cesc Chunseong Park, Byeongchang Kim, Gunhee Kim; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 895-903
We address personalization issues of image captioning, which have not been discu ssed yet in previous research. For a query image, we aim to generate a descripti ve sentence, accounting for prior knowledge such as the user's active vocabulari es in previous documents. As applications of personalized image captioning, we t ackle two post automation tasks: hashtag prediction and post generation, on our newly collected Instagram dataset, consisting of 1.1M posts from 6.3K users. We propose a novel captioning model named Context Sequence Memory Network (CSMN). I ts unique updates over previous memory network models include (i) exploiting mem ory as a repository for multiple types of context information, (ii) appending pr eviously generated words into memory to capture long-term information without su ffering from the vanishing gradient problem, and (iii) adopting CNN memory struc ture to jointly represent nearby ordered memory slots for better context underst anding. With quantitative evaluation and user studies via Amazon Mechanical Turk , we show the effectiveness of the three novel features of CSMN and its performa nce enhancement for personalized image captioning over state-of-the-art captioni ng models.
**********

Scalable Surface Reconstruction From Point Clouds With Extreme Scale and Density Diversity
Christian Mostegel, Rudolf Prettenthaler, Friedrich Fraundorfer, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 904-913
In this paper we present a scalable approach for robustly computing a 3D surface mesh from multi-scale multi-view stereo point clouds that can handle extreme jumps of point density (in our experiments three orders of magnitude). The backbone of our approach is a combination of octree data partitioning, local Delaunay tetrahedralization and graph cut optimization. Graph cut optimization is used twice, once to extract surface hypotheses from local Delaunay tetrahedralizations and once to merge overlapping surface hypotheses even when the local tetrahedralizations do not share the same topology. This formulation allows us to obtain a constant memory consumption per sub-problem while at the same time retaining the density independent interpolation properties of the Delaunay-based optimization. On multiple public datasets, we demonstrate that our approach is highly competitive with the state-of-the-art in terms of accuracy, completeness and outlier resilience. Further, we demonstrate the multi-scale potential of our approach by processing a newly recorded dataset with 2 billion points and a point density variation of more than four orders of magnitude - requiring less than 9GB of RAM per process.
**********
Weakly Supervised Cascaded Convolutional Networks
Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 914-922
Object detection is a challenging task in visual understanding domain, and even more so if the supervision is to be weak. Recently, few efforts to handle the task without expensive human annotations is established by promising deep neural network. A new architecture of cascaded networks is proposed to learn a convolutional neural network (CNN) under such conditions. We introduce two such architectures, with either two cascade stages or three which are trained in an end-to-end pipeline. The first stage of both architectures extracts best candidate of class specific region proposals by training a fully convolutional network. In the case of the three stage architecture, the middle stage provides object segmentation, using the output of the activation maps of first stage. The final stage of both architectures is a part of a convolutional neural network that performs multiple instance learning on proposals extracted in the previous stage(s). Our experiments on the PASCAL VOC 2007, 2010, 2012 and large scale object datasets, ILSVRC 2013, 2014 datasets show improvements in the areas of weakly-supervised object detection, classification and localization.
**********
Exclusivity-Consistency Regularized Multi-View Subspace Clustering
Xiaobo Wang, Xiaojie Guo, Zhen Lei, Changqing Zhang, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 923-931
Multi-view subspace clustering aims to partition a set of multi-source data into their underlying groups. To boost the performance of multi-view clustering, numerous subspace learning algorithms have been developed in recent years, but with rare exploitation of the representation complementarity between different views as well as the indicator consistency among the representations, let alone considering them simultaneously. In this paper, we propose a novel multi-view subspace clustering model that attempts to harness the complementary information between different representations by introducing a novel position-aware exclusivity term. Meanwhile, a consistency term is employed to make these complementary representations to further have a common indicator. We formulate the above concerns into a unified optimization framework. Experimental results on several benchmark datasets are conducted to reveal the effectiveness of our algorithm over other state-of-the-arts.
**********

Look Into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing
Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 932-940
Human parsing has recently attracted a lot of research interests due to its huge application potentials. However existing datasets have limited number of images and annotations, and lack the variety of human appearances and the coverage of challenging cases in unconstrained environment. In this paper, we introduce a new benchmark "Look into Person (LIP)" that makes a significant advance in terms of scalability, diversity and difficulty, a contribution that we feel is crucial for future developments in human-centric analysis. This comprehensive dataset contains over 50,000 elaborately annotated images with 19 semantic part labels, which are captured from a wider range of viewpoints, occlusions and background complexity. Given these rich annotations we perform detailed analysis of the leading human parsing approaches, gaining insights into the success and failures of these methods. Furthermore, in contrast to the existing efforts on improving the feature discriminative capability, we solve human parsing by exploring a novel self-supervised structure-sensitive learning approach, which imposes human pose structures into parsing results without resorting to extra supervision (i.e., no need for specifically labeling human joints in model training). Our self-supervised learning framework can be injected into any advanced neural networks to help incorporate rich high-level knowledge regarding human joints from a global perspective and improve the parsing results. Extensive evaluations on our LIP and the public PASCAL-Person-Part dataset demonstrate the superiority of our method.
**********
Semi-Calibrated Near Field Photometric Stereo
Fotios Logothetis, Roberto Mecca, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 941-950
3D reconstruction from shading information through Photometric Stereo is considered a very challenging problem in Computer Vision. Although this technique can potentially provide highly detailed shape recovery, its accuracy is critically dependent on a numerous set of factors among them the reliability of the light sources in emitting a constant amount of light. In this work, we propose a novel variational approach to solve the so called semi-calibrated near field Photometric Stereo problem, where the positions but not the brightness of the light sources are known. Additionally, we take into account realistic modeling features such as perspective viewing geometry and heterogeneous scene composition, containing both diffuse and specular objects. Furthermore, we also relax the point light source assumption that usually constraints the near field formulation by explicitly calculating the light attenuation maps. Synthetic experiments are performed for quantitative evaluation for a wide range of cases whilst real experiments provide comparisons, qualitatively outperforming the state of the art.
**********
Finding Tiny Faces
Peiyun Hu, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 951-959
Though tremendous strides have been made in object recognition, one of the remaining open challenges is detecting small objects. We explore three aspects of the problem in the context of finding small faces: the role of scale invariance, image resolution, and contextual reasoning. While most recognition approaches aim to be scale-invariant, the cues for recognizing a 3px tall face are fundamentally different than those for recognizing a 300px tall face. We take a different approach and train separate detectors for different scales. To maintain efficiency, detectors are trained in a multi-task fashion: they make use of features extracted from multiple layers of single (deep) feature hierarchy. While training detectors for large objects is straightforward, the crucial challenge remains training detectors for small objects. We show that context is crucial, and define templates that make use of massively-large receptive fields (where 99% of the template extends beyond the object of interest). Finally, we explore the role of scal

e in pre-trained deep networks, providing ways to extrapolate networks tuned for limited scales to rather extreme ranges. We demonstrate state-of-the-art results on massively-benchmarked face datasets (FDDB and WIDER FACE). In particular, when compared to prior art on WIDER FACE, our results reduce error by a factor of 2 (our models produce an AP of 82% while prior art ranges from 29-64%).
**********

Visual-Inertial-Semantic Scene Representation for 3D Object Detection
Jingming Dong, Xiaohan Fei, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 960-970
We describe a system to detect objects in three-dimensional space using video and inertial sensors (accelerometer and gyrometer), ubiquitous in modern mobile platforms from phones to drones. Inertials afford the ability to impose class-specific scale priors for objects, and provide a global orientation reference. A minimal sufficient representation, the posterior of semantic (identity) and syntactic (pose) attributes of objects in space, can be decomposed into a geometric term, which can be maintained by a localization-and-mapping filter, and a likelihood function, which can be approximated by a discriminatively-trained convolutional neural network The resulting system can process the video stream causally in real time, and provides a representation of objects in the scene that is persistent: Confidence in the presence of objects grows with evidence, and objects previously seen are kept in memory even when temporarily occluded, with their return into view automatically predicted to prime re-detection.
**********

ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification
Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, Bryan Russell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 971-980
In this work, we introduce a new video representation for action classification that aggregates local convolutional features across the entire spatio-temporal extent of the video. We do so by integrating state-of-the-art two-stream networks with learnable spatio-temporal feature aggregation. The resulting architecture is end-to-end trainable for whole-video classification. We investigate different strategies for pooling across space and time and combining signals from the different streams. We find that: (i) it is important to pool jointly across space and time, but (ii) appearance and motion streams are best aggregated into their own separate representations. Finally, we show that our representation outperforms the two-stream base architecture by a large margin (13% relative) as well as outperforms other baselines with comparable base architectures on HMDB51, UCF101, and Charades video classification benchmarks.
**********

Predictive-Corrective Networks for Action Detection
Achal Dave, Olga Russakovsky, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 981-990
While deep feature learning has revolutionized techniques for static-image understanding, the same does not quite hold for video processing. Architectures and optimization techniques used for video are largely based off those for static images, potentially underutilizing rich video information. In this work, we rethink both the underlying network architecture and the stochastic learning paradigm for temporal data. To do so, we draw inspiration from classic theory on linear dynamic systems for modeling time series. By extending such models to include nonlinear mappings, we derive a series of novel recurrent neural networks that sequentially make top-down predictions about the future and then correct those predictions with bottom-up observations. Predictive-corrective networks have a number of desirable properties: (1) they can adaptively focus computation on "surprising" frames where predictions require large corrections, (2) they simplify learning in that only "residual-like" corrective terms need to be learned over time and (3) they naturally decorrelate an input data stream in a hierarchical fashion, producing a more reliable signal for learning at each layer of a network. We provide an extensive analysis of our lightweight and interpretable framework, and demonstrate that our model is competitive with the two-stream network on three ch

allenging datasets without the need for computationally expensive optical flow.
**********

FastMask: Segment Multi-Scale Object Candidates in One Shot
Hexiang Hu, Shiyi Lan, Yuning Jiang, Zhimin Cao, Fei Sha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 991-999
Objects appear to scale differently in natural images. This fact requires methods dealing with object-centric tasks (e.g. object proposal) to have robust performance over variances in object scales. In the paper, we present a novel segment proposal framework, namely FastMask, which takes advantage of hierarchical features in deep convolutional neural networks to segment multi-scale objects in one shot. Innovatively, we adapt segment proposal network into three different functional components (body, neck and head). We further propose a weight-shared residual neck module as well as a scale-tolerant attentional head module for efficient one-shot inference. On MS COCO benchmark, the proposed FastMask outperforms all state-of-the-art segment proposal methods in average recall being 2 5 times faster. Moreover, with a slight trade-off in accuracy, FastMask can segment objects in near real time ( 13 fps) with 800*600 resolution images, demonstrating its potential in practical applications. Our implementation is available on https://github.com/voidrank/FastMask.
**********

A Combinatorial Solution to Non-Rigid 3D Shape-To-Image Matching
Florian Bernard, Frank R. Schmidt, Johan Thunberg, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1000-1009
We propose a combinatorial solution for the problem of non-rigidly matching a 3D shape to 3D image data. To this end, we model the shape as a triangular mesh and allow each triangle of this mesh to be rigidly transformed to achieve a suitable matching to the image. By penalising the distance and the relative rotation between neighbouring triangles our matching compromises between the image and the shape information. In this paper, we resolve two major challenges: Firstly, we address the resulting large and NP-hard combinatorial problem with a suitable graph-theoretic approach. Secondly, we propose an efficient discretisation of the unbounded 6-dimensional Lie group SE(3). To our knowledge this is the first combinatorial formulation for non-rigid 3D shape-to-image matching. In contrast to existing local (gradient descent) optimisation methods, we obtain solutions that do not require a good initialisation and that are within a bound of the optimal solution. We evaluate the proposed combinatorial method on the two problems of non-rigid 3D shape-to-shape and non-rigid 3D shape-to-image registration and demonstrate that it provides promising results.
**********

Interpretable Structure-Evolving LSTM
Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, Eric P. Xing ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1010-1019
This paper develops a general framework for learning interpretable data representation via Long Short-Term Memory (LSTM) recurrent neural networks over hierarchical graph structures. Instead of learning LSTM models over the pre-fixed structures, we propose to further learn the intermediate interpretable multi-level graph structures in a progressive and stochastic way from data during the LSTM network optimization. We thus call this model the structure-evolving LSTM. In particular, starting with an initial element-level graph representation where each node is a small data element, the structure-evolving LSTM gradually evolves the multi-level graph representations by stochastically merging the graph nodes with high compatibilities along the stacked LSTM layers. In each LSTM layer, we estimate the compatibility of two connected nodes from their corresponding LSTM gate outputs, which is used to generate a merging probability. The candidate graph structures are accordingly generated where the nodes are grouped into cliques with their merging probabilities. We then produce the new graph structure with a Metropolis-Hasting algorithm, which alleviates the risk of getting stuck in local optim

ums by stochastic sampling with an acceptance probability. Once a graph structur e is accepted, a higher-level graph is then constructed by taking the partitione d cliques as its nodes. During the evolving process, representation becomes more abstracted in higher-levels where redundant information is filtered out, allowi ng more efficient propagation of long-range data dependencies. We evaluate the e ffectiveness of structure-evolving LSTM in the application of semantic object pa rsing and demonstrate its advantage over state-of-the-art LSTM models on standar d benchmarks.

**********

Generating the Future With Adversarial Transformers

Carl Vondrick, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1020-1028

We learn models to generate the immediate future in video. This problem has two main challenges. Firstly, since the future is uncertain, models should be multi-modal, which can be difficult to learn. Secondly, since the future is similar to the past, models store low-level details, which complicates learning of high-le vel semantics. We propose a framework to tackle both of these challenges. We pre sent a model that generates the future by transforming pixels in the past. Our a pproach explicitly disentangles the model's memory from the prediction, which he lps the model learn desirable invariances. Experiments suggest that this model c an generate short videos of plausible futures. We believe predictive models have many applications in robotics, health-care, and video understanding.

**********

Budget-Aware Deep Semantic Video Segmentation

Behrooz Mahasseni, Sinisa Todorovic, Alan Fern; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1029-1038

In this work, we study a poorly understood trade-off between accuracy and runtim e costs for deep semantic video segmentation. While recent work has demonstrated advantages of learning to speed-up deep activity detection, it is not clear if similar advantages will hold for our very different segmentation loss function, which is defined over individual pixels across the frames. In deep video segment ation, the most time consuming step represents the application of a CNN to every frame for assigning class labels to every pixel, typically taking 6-9 times of the video footage. This motivates our new budget-aware framework that learns to optimally select a small subset of frames for pixelwise labeling by a CNN, and t hen efficiently interpolates the obtained segmentations to yet unprocessed frame s. This interpolation may use either a simple optical-flow guided mapping of pix el labels, or another significantly less complex and thus faster CNN. We formali ze the frame selection as a Markov Decision Process, and specify a Long Short-Te rm Memory (LSTM) network to model a policy for selecting the frames. For trainin g the LSTM, we develop a policy-gradient reinforcement-learning approach for app roximating the gradient of our non-decomposable and non-differentiable objective . Evaluation on two benchmark video datasets show that our new framework is able to significantly reduce computation time, and maintain competitive video segmen tation accuracy under varying budgets.

**********

Spatially Adaptive Computation Time for Residual Networks

Michael Figurnov, Maxwell D. Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitr y Vetrov, Ruslan Salakhutdinov; Proceedings of the IEEE Conference on Computer V ision and Pattern Recognition (CVPR), 2017, pp. 1039-1048

This paper proposes a deep learning architecture based on Residual Network that dynamically adjusts the number of executed layers for the regions of the image. This architecture is end-to-end trainable, deterministic and problem-agnostic. I t is therefore applicable without any modifications to a wide range of computer vision problems such as image classification, object detection and image segment ation. We present experimental results showing that this model improves the comp utational efficiency of Residual Networks on the challenging ImageNet classifica tion and COCO object detection datasets. Additionally, we evaluate the computati on time maps on the visual saliency dataset cat2000 and find that they correlate surprisingly well with human eye fixation positions.

**********
Order-Preserving Wasserstein Distance for Sequence Matching
Bing Su, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1049-1057
We present a new distance measure between sequences that can tackle local temporal distortion and periodic sequences with arbitrary starting points. Through viewing the instances of sequences as empirical samples of an unknown distribution, we cast the calculation of the distance between sequences as the optimal transport problem. To preserve the inherent temporal relationships of the instances in sequences, we smooth the optimal transport problem with two novel temporal regularization terms. The inverse difference moment regularization enforces transport with local homogeneous structures, and the KL-divergence with a prior distribution regularization prevents transport between instances with far temporal positions. We show that this problem can be efficiently optimized through the matrix scaling algorithm. Extensive experiments on different datasets with different classifiers show that the proposed distance outperforms the traditional DTW variants and the smoothed optimal transport distance without temporal regularization.
**********
Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction
Richard Zhang, Phillip Isola, Alexei A. Efros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1058-1067
We propose split-brain autoencoders, a straightforward modification of the traditional autoencoder architecture, for unsupervised representation learning. The method adds a split to the network, resulting in two disjoint sub-networks. Each sub-network is trained to perform a difficult task -- predicting one subset of the data channels from another. Together, the sub-networks extract features from the entire input signal. By forcing the network to solve cross-channel prediction tasks, we induce a representation within the network which transfers well to other, unseen tasks. This method achieves state-of-the-art performance on several large-scale transfer learning benchmarks.
**********
SRN: Side-output Residual Network for Object Symmetry Detection in the Wild
Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, Qixiang Ye; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1068-1076
In this paper, we establish a baseline for object symmetry detection in complex backgrounds by presenting a new benchmark and an end-to-end deep learning approach, opening up a promising direction for symmetry detection in the wild. The new benchmark, named Sym-PASCAL, spans challenges including object diversity, multi-objects, part-invisibility, and various complex backgrounds that are far beyond those in existing datasets. The proposed symmetry detection approach, named Side-output Residual Network (SRN), leverages output Residual Units (RUs) to fit the errors between the object symmetry ground-truth and the outputs of RUs. By stacking RUs in a deep-to-shallow manner, SRN exploits the 'flow' of errors among multiple scales to ease the problems of fitting complex outputs with limited layers, suppressing the complex backgrounds, and effectively matching object symmetry of different scales. Experimental results validate both the benchmark and its challenging aspects related to real-world images, and the state-of-the-art performance of our symmetry detection approach. The benchmark and the code for SRN are publicly available at https://github.com/KevinKecc/SRN.
**********
Spindle Net: Person Re-Identification With Human Body Region Guided Feature Decomposition and Fusion
Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1077-1085
Person re-identification (ReID) is an important task in video surveillance and has various applications. It is non-trivial due to complex background clutters, varying illumination conditions, and uncontrollable camera settings. Moreover, the person body misalignment caused by detectors or pose variations is sometimes t

oo severe for feature matching across images. In this study, we propose a novel Convolutional Neural Network (CNN), called Spindle Net, based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion. It is the first time human body structure information is considered in a CNN framework to facilitate feature learning. The proposed Spindle Net brings unique advantages: 1) it separately captures semantic features from different body regions thus the macro- and micro-body features can be well aligned across images, 2) the learned region features from different semantic regions are merged with a competitive scheme and discriminative features can be well preserved. State of the art performance can be achieved on multiple datasets by large margins. We further demonstrate the robustness and effectiveness of the proposed Spindle Net on our proposed dataset SenseReID without fine-tuning.

**********

Borrowing Treasures From the Wealthy: Deep Transfer Learning Through Selective Joint Fine-Tuning

Weifeng Ge, Yizhou Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1086-1095

Deep neural networks require a large amount of labeled training data during supervised learning. However, collecting and labeling so much data might be infeasible in many cases. In this paper, we introduce a deep transfer learning scheme, called selective joint fine-tuning, for improving the performance of deep learning tasks with insufficient training data. In this scheme, a target learning task with insufficient training data is carried out simultaneously with another source learning task with abundant training data. However, the source learning task does not use all existing training data. Our core idea is to identify and use a subset of training images from the original source learning task whose low-level characteristics are similar to those from the target learning task, and jointly fine-tune shared convolutional layers for both tasks. Specifically, we compute descriptors from linear or nonlinear filter bank responses on training images from both tasks, and use such descriptors to search for a desired subset of training samples for the source learning task. Experiments demonstrate that our deep transfer learning scheme achieves state-of-the-art performance on multiple visual classification tasks with insufficient training data for deep learning. Such tasks include Caltech 256, MIT Indoor 67, and fine-grained classification problems (Oxford Flowers 102 and Stanford Dogs 120). In comparison to fine-tuning without a source domain, the proposed method can improve the classification accuracy by 2% - 10% using a single model. Codes and models are available at https://github.com/ZYYSzj/Selective-Joint-Fine-tuning.

**********

Unified Embedding and Metric Learning for Zero-Exemplar Event Detection

Noureldien Hussein, Efstratios Gavves, Arnold W.M. Smeulders; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1096-1105

Event detection in unconstrained videos is conceived as a content-based video retrieval with two modalities: textual and visual. Given a text describing a novel event, the goal is to rank related videos accordingly. This task is zero-exemplar, no video examples are given to the novel event. Related works train a bank of concept detectors on external data sources. These detectors predict confidence scores for test videos, which are ranked and retrieved accordingly. In contrast, we learn a joint space in which the visual and textual representations are embedded. The space casts a novel event as a probability of pre-defined events. Also, it learns to measure the distance between an event and its related videos. Our model is trained end-to-end on publicly available EventNet. When applied to TRECVID Multimedia Event Detection dataset, it outperforms the state-of-the-art by a considerable margin.

**********

A Practical Method for Fully Automatic Intrinsic Camera Calibration Using Directionally Encoded Light

Mahdi Abbaspour Tehrani, Thabo Beeler, Anselm Grundhofer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1106-1

114
Calibrating the intrinsic properties of a camera is one of the fundamental tasks required for a variety of computer vision and image processing tasks. The precise measurement of focal length, location of the principal point as well as distortion parameters of the lens is crucial, for example, for 3D reconstruction. Although a variety of methods exist to achieve this goal, they are often cumbersome to carry out, require substantial manual interaction, expert knowledge, and a significant operating volume. We propose a novel calibration method based on the usage of directionally encoded light rays for estimating the intrinsic parameters. It enables a fully automatic calibration with a small device mounted close to the front lens element and still enables an accuracy comparable to standard methods even when the lens is focused up to infinity. Our method overcomes the mentioned limitations since it guarantees an accurate calibration without any human intervention while requiring only a limited amount of space. Besides that, the approach also allows to estimate the distance of the focal plane as well as the size of the aperture. We demonstrate the advantages of the proposed method by evaluating several camera/lens configurations using prototypical devices.
**********

Modeling Relationships in Referential Expressions With Compositional Modular Networks
Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1115-1124
People often refer to entities in an image in terms of their relationships with other entities. For example, "the black cat sitting under the table" refers to both a "black cat" entity and its relationship with another "table" entity. Understanding these relationships is essential for interpreting and grounding such natural language expressions. Most prior work focuses on either grounding entire referential expressions holistically to one region, or localizing relationships based on a fixed set of categories. In this paper we instead present a modular deep architecture capable of analyzing referential expressions into their component parts, identifying entities and relationships mentioned in the input expression and grounding them all in the scene. We call this approach Compositional Modular Networks (CMNs): a novel architecture that learns linguistic analysis and visual inference end-to-end. Our approach is built around two types of neural modules that inspect local regions and pairwise interactions between regions. We evaluate CMNs on multiple referential expression datasets, outperforming state-of-the-art approaches on all tasks.
**********

Image-To-Image Translation With Conditional Adversarial Networks
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125-1134
We investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. We demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. Moreover, since the release of the pix2pix software associated with this paper, hundreds of twitter users have posted their own artistic experiments using our system. As a community, we no longer hand-engineer our mapping functions, and this work suggests we can achieve reasonable results without handengineering our loss functions either.
**********

Counting Everyday Objects in Everyday Scenes
Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1135-1144

We are interested in counting the number of instances of object classes in natur
al, everyday images. Previous counting approaches tackle the problem in restrict
ed domains such as counting pedestrians in surveillance videos. Counts can also
be estimated from outputs of other vision tasks like object detection. In this w
ork, we build dedicated models for counting designed to tackle the large varianc
e in counts, appearances, and scales of objects found in natural scenes. Our app
roach is inspired by the phenomenon of subitizing - the ability of humans to mak
e quick assessments of counts given a perceptual signal, for small count values.
 Given a natural scene, we employ a divide and conquer strategy while incorporat
ing context across the scene to adapt the subitizing idea to counting. Our appro
ach offers consistent improvements over numerous baseline approaches for countin
g on the PASCAL VOC 2007 and COCO datasets. Subsequently, we study how counting
can be used to improve object detection. We then show a proof of concept applica
tion of our counting methods to the task of Visual Question Answering, by studyi
ng the 'how many?' questions in the VQA and COCO-QA datasets.
**********

Hand Keypoint Detection in Single Images Using Multiview Bootstrapping
Tomas Simon, Hanbyul Joo, Iain Matthews, Yaser Sheikh; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1145-1153
We present an approach that uses a multi-camera system to train fine-grained det
ectors for keypoints that are prone to occlusion, such as the joints of a hand.
We call this procedure multiview bootstrapping: first, an initial keypoint detec
tor is used to produce noisy labels in multiple views of the hand. The noisy det
ections are then triangulated in 3D using multiview geometry or marked as outlie
rs. Finally, the reprojected triangulations are used as new labeled training dat
a to improve the detector. We repeat this process, generating more labeled data
in each iteration. We derive a result analytically relating the minimum number o
f views to achieve target true and false positive rates for a given detector. Th
e method is used to train a hand keypoint detector for single images. The result
ing keypoint detector runs in realtime on RGB images and has accuracy comparable
 to methods that use depth sensors. The single view detector, triangulated over
multiple views, enables 3D markerless hand motion capture with complex object in
teractions.
**********

Knowledge Acquisition for Visual Question Answering via Iterative Querying
Yuke Zhu, Joseph J. Lim, Li Fei-Fei; Proceedings of the IEEE Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2017, pp. 1154-1163
Humans possess an extraordinary ability to learn new skills and new knowledge fo
r problem solving. Such learning ability is also required by an automatic model
to deal with arbitrary, open-ended questions in the visual world. We propose a n
eural-based approach to acquiring task- driven information for visual question a
nswering (VQA). Our model proposes queries to actively acquire relevant informat
ion from external auxiliary data. Supporting evidence from either human-curated
or automatic sources is encoded and stored into a memory bank. We show that acqu
iring task-driven evidence effectively improves model performance on both the Vi
sual7W and VQA datasets; moreover, these queries offer certain level of interpre
tability in our iterative QA model.
**********

Few-Shot Object Recognition From Machine-Labeled Web Images
Zhongwen Xu, Linchao Zhu, Yi Yang; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2017, pp. 1164-1172
With the tremendous advances made by Convolutional Neural Networks (ConvNets) on
 object recognition, we can now easily obtain adequately reliable machine-labele
d annotations easily from predictions by off-the-shelf ConvNets. In this work, w
e present an "abstraction memory" based framework for few-shot learning, buildin
g upon machine-labeled image annotations. Our method takes large-scale machine-a
nnotated dataset (e.g., OpenImages) as an external memory bank. In the external
memory bank, the information is stored in the memory slots in the form of key-va
lue, in which image feature is regarded as the key and the label embedding serve
s as the value. When queried by the few-shot examples, our model selects visuall

y similar data from the external memory bank and writes the useful information o
btained from related external data into another memory bank, i.e., abstraction m
emory. Long Short-Term Memory (LSTM) controllers and attention mechanisms are ut
ilized to guarantee the data written to the abstraction memory correlates with t
he query example. The abstraction memory concentrates information from the exter
nal memory bank to make the few-shot recognition effective. In the experiments,
we first confirm that our model can learn to conduct few-shot object recognition
 on clean human-labeled data from the ImageNet dataset. Then, we demonstrate tha
t with our model, machine-labeled image annotations are very effective and abund
ant resources for performing object recognition on novel categories. Experimenta
l results show that our proposed model with machine-labeled annotations achieves
 great results, with only a 1% difference in accuracy between the machine-labele
d annotations and the human-labeled annotations.
**********

The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Qu
estions

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1173-1182

One of the most intriguing features of the Visual Question Answering (VQA) chall
enge is the unpredictability of the questions. Extracting the information requir
ed to answer them demands a variety of image operations from detection and count
ing, to segmentation and reconstruction. To train a method to perform even one o
f these operations accurately from image,question,answer tuples would be challen
ging, but to aim to achieve them all with a limited set of such training data se
ems ambitious at best. Our method thus learns how to exploit a set of external o
ff-the-shelf algorithms to achieve its goal, an approach that has something in c
ommon with the Neural Turing Machine. The core of our proposed method is a new c
o-attention model. In addition, the proposed approach generates human-readable r
easons for its decision, and can still be trained end-to-end without ground trut
h reasons being given. We demonstrate the effectiveness on two publicly availabl
e datasets, Visual Genome and VQA, and show that it produces the state-of-the-ar
t results in both cases.
**********

Learning Deep Binary Descriptor With Multi-Quantization

Yueqi Duan, Jiwen Lu, Ziwei Wang, Jianjiang Feng, Jie Zhou; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1183
-1192

In this paper, we propose an unsupervised feature learning method called deep bi
nary descriptor with multi-quantization (DBD-MQ) for visual matching. Existing l
earning-based binary descriptors such as compact binary face descriptor (CBFD) a
nd DeepBit utilize the rigid sign function for binarization despite of data dist
ributions, thereby suffering from severe quantization loss. In order to address
the limitation, our DBD-MQ considers the binarization as a multi-quantization ta
sk. Specifically, we apply a K-AutoEncoders (KAEs) network to jointly learn the
parameters and the binarization functions under a deep learning framework, so th
at discriminative binary descriptors can be obtained with a fine-grained multi-q
uantization. Extensive experimental results on different visual analysis includi
ng patch retrieval, image matching and image retrieval show that our DBD-MQ outp
erforms most existing binary feature descriptors.
**********

Joint Discriminative Bayesian Dictionary and Classifier Learning

Naveed Akhtar, Ajmal Mian, Fatih Porikli; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1193-1202

We propose to jointly learn a Discriminative Bayesian dictionary along a linear
classifier using coupled Beta-Bernoulli Processes. Our representation model uses
 separate base measures for the dictionary and the classifier, but associates th
em to the class-specific training data using the same Bernoulli distributions. T
he Bernoulli distributions control the frequency with which the factors (e.g. di
ctionary atoms) are used in data representations, and they are inferred while ac
counting for the class labels in our approach. To further encourage discriminati

on in the dictionary, our model uses separate (sets of) Bernoulli distributions to represent data from different classes. Our approach adaptively learns the association between the dictionary atoms and the class labels while tailoring the classifier to this relation with a joint inference over the dictionary and the classifier. Once a test sample is represented over the dictionary, its representation is accurately labelled by the classifier due to the strong coupling between the dictionary and the classifier. We derive the Gibbs Sampling equations for our joint representation model and test our approach for face, object, scene and action recognition to establish its effectiveness.
**********

A Graph Regularized Deep Neural Network for Unsupervised Image Representation Learning
Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, Qingming Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1203-1211
Deep Auto-Encoder (DAE) has shown its promising power in high-level representation learning. From the perspective of manifold learning, we propose a graph regularized deep neural network (GR-DNN) to endue traditional DAEs with the ability of retaining local geometric structure. A deep-structured regularizer is formulated upon multi-layer perceptions to capture this structure. The robust and discriminative embedding space is learned to simultaneously preserve the high-level semantics and the geometric structure within local manifold tangent space. Theoretical analysis presents the close relationship between the proposed graph regularizer and the graph Laplacian regularizer in terms of the optimization objective. We also alleviate the growth of the network complexity by introducing the anchor-based bipartite graph, which guarantees the good scalability for large scale data. The experiments on four datasets show the comparable results of the proposed GR-DNN with the state-of-the-art methods.
**********

HSfM: Hybrid Structure-from-Motion
Hainan Cui, Xiang Gao, Shuhan Shen, Zhanyi Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1212-1221
Structure-from-Motion (SfM) methods can be broadly categorized as incremental or global according to their ways to estimate initial camera poses. While incremental system has advanced in robustness and accuracy, the efficiency remains its key challenge. To solve this problem, global reconstruction system simultaneously estimates all camera poses from the epipolar geometry graph, but it is usually sensitive to outliers. In this work, we propose a new hybrid SfM method to tackle the issues of efficiency, accuracy and robustness in a unified framework. More specifically, we propose an adaptive community-based rotation averaging method first to estimate camera rotations in a global manner. Then, based on these estimated camera rotations, camera centers are computed in an incremental way. Extensive experiments show that our hybrid method performs similarly or better than many of the state-of-the-art global SfM approaches, in terms of computational efficiency, while achieves similar reconstruction accuracy and robustness with two other state-of-the-art incremental SfM approaches.
**********

Perceptual Generative Adversarial Networks for Small Object Detection
Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1222-1230
Detecting small objects is notoriously challenging due to their low resolution and noisy representation. Existing object detection pipelines usually detect small objects through learning representations of all the objects at multiple scales. However, the performance gain of such ad hoc architectures is usually limited to pay off the computational cost. In this work, we address the small object detection problem by developing a single architecture that internally lifts representations of small objects to super-resolved ones, achieving similar characteristics as large objects and thus more discriminative for detection. For this purpose, we propose a new Perceptual Generative Adversarial Network (Perceptual GAN) m

odel that improves small object detection through narrowing representation diffe
rence of small objects from the large ones. Specifically, its generator learns t
o transfer perceived poor representations of the small objects to super-resolved
 ones that are similar enough to real large objects to fool a competing discrimi
nator. Meanwhile its discriminator competes with the generator to identify the g
enerated representation and imposes an additional perceptual requirement - gener
ated representations of small objects must be beneficial for detection purpose -
 on the generator. Extensive evaluations on the challenging Tsinghua-Tencent 100
K and the Caltech benchmark well demonstrate the superiority of Perceptual GAN i
n detecting small objects, including traffic signs and pedestrians, over well-es
tablished state-of-the-arts.
**********

Deep Roots: Improving CNN Efficiency With Hierarchical Filter Groups
Yani Ioannou, Duncan Robertson, Roberto Cipolla, Antonio Criminisi; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,
pp. 1231-1240
We propose a new method for creating computationally efficient and compact convo
lutional neural networks (CNNs) using a novel sparse connection structure that r
esembles a tree root. This allows a significant reduction in computational cost
and number of parameters compared to state-of-the-art deep CNNs, without comprom
ising accuracy, by exploiting the sparsity of inter-layer filter dependencies. W
e validate our approach by using it to train more efficient variants of state-of
-the-art CNN architectures, evaluated on the CIFAR10 and ILSVRC datasets. Our re
sults show similar or higher accuracy than the baseline architectures with much
less computation, as measured by CPU and GPU timings. For example, for ResNet 50
, our model has 40% fewer parameters, 45% fewer floating point operations, and i
s 31% (12%) faster on a CPU (GPU). For the deeper ResNet 200 our model has 48% f
ewer parameters and 27% fewer floating point operations, while maintaining state
-of-the-art accuracy. For GoogLeNet, our model has 7% fewer parameters and is 21
% (16%) faster on a CPU (GPU).
**********

Anti-Glare: Tightly Constrained Optimization for Eyeglass Reflection Removal
Tushar Sandhan, Jin Young Choi; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2017, pp. 1241-1250
Absence of a clear eye visibility not only degrades the aesthetic value of an en
tire face image but also creates difficulties in many computer vision tasks. Eve
n mild reflections produce the undesired superpositions of visual information, w
hose decomposition into the background and reflection layers using a single imag
e is a highly ill-posed problem. In this work, we enforce the tight constraints
derived by thoroughly analysing the properties of an eyeglass reflection. In add
ition, our strategy regularizes gradients of the reflection layer to be highly s
parse and proposes the facial symmetry prior via formulating a non-convex optimi
zation scheme, which removes the reflections within a few iterations. Experiment
s on frontal face image inputs demonstrate the high quality reflection removal r
esults and improvement of the iris detection rate.
**********

Xception: Deep Learning With Depthwise Separable Convolutions
Francois Chollet; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2017, pp. 1251-1258
We present an interpretation of Inception modules in convolutional neural networ
ks as being an intermediate step in-between regular convolution and the depthwis
e separable convolution operation (a depthwise convolution followed by a pointwi
se convolution). In this light, a depthwise separable convolution can be underst
ood as an Inception module with a maximally large number of towers. This observa
tion leads us to propose a novel deep convolutional neural network architecture
inspired by Inception, where Inception modules have been replaced with depthwise
 separable convolutions. We show that this architecture, dubbed Xception, slight
ly outperforms Inception V3 on the ImageNet dataset (which Inception V3 was desi
gned for), and significantly outperforms Inception V3 on a larger image classifi
cation dataset comprising 350 million images and 17,000 classes. Since the Xcept

ion architecture has the same number of parameters as Inception V3, the performa nce gains are not due to increased capacity but rather to a more efficient use o f model parameters.
**********

Learning Detailed Face Reconstruction From a Single Image
Elad Richardson, Matan Sela, Roy Or-El, Ron Kimmel; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1259-1268
Reconstructing the detailed geometric structure of a face from a given image is a key to many computer vision and graphics applications, such as motion capture and reenactment. The reconstruction task is challenging as human faces vary exte nsively when considering expressions, poses, textures, and intrinsic geometries. While many approaches tackle this complexity by using additional data to recons truct the face of a single subject, extracting facial surface from a single imag e remains a difficult problem. As a result, single-image based methods can usual ly provide only a rough estimate of the facial geometry. In contrast, we propose to leverage the power of convolutional neural networks to produce a highly deta iled face reconstruction from a single image. For this purpose, we introduce an end-to-end CNN framework which derives the shape in a coarse-to-fine fashion. Th e proposed architecture is composed of two main blocks, a network that recovers the coarse facial geometry (CoarseNet), followed by a CNN that refines the facia l features of that geometry (FineNet). The proposed networks are connected by a novel layer which renders a depth image given a mesh in 3D. Unlike object recogn ition and detection problems, there are no suitable datasets for training CNNs t o perform face geometry reconstruction. Therefore, our training regime begins wi th a supervised phase, based on synthetic images, followed by an unsupervised ph ase that uses only unconstrained facial images. The accuracy and robustness of t he proposed model is demonstrated by both qualitative and quantitative evaluatio n tests.
**********

Stereo-Based 3D Reconstruction of Dynamic Fluid Surfaces by Global Optimization
Yiming Qian, Minglun Gong, Yee-Hong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1269-1278
3D reconstruction of dynamic fluid surfaces is an open and challenging problem i n computer vision. Unlike previous approaches that reconstruct each surface poin t independently and often return noisy depth maps, we propose a novel global opt imization-based approach that recovers both depths and normals of all 3D points simultaneously. Using the traditional refraction stereo setup, we capture the wa vy appearance of a pre-generated random pattern, and then estimate the correspon dences between the captured images and the known background by tracking the patt ern. Assuming that the light is refracted only once through the fluid interface, we minimize an objective function that incorporates both the cross-view normal consistency constraint and the single-view normal consistency constraints. The k ey idea is that the normals required for light refraction based on Snell's law f rom one view should agree with not only the ones from the second view, but also the ones estimated from local 3D geometry. Moreover, an effective reconstruction error metric is designed for estimating the refractive index of the fluid. We r eport experimental results on both synthetic and real data demonstrating that th e proposed approach is accurate and shows superiority over the conventional ster eo-based method.
**********

Deep Video Deblurring for Hand-Held Cameras
Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, Oliver Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition (CVPR), 2017, pp. 1279-1288
Motion blur from camera shake is a major problem in videos captured by hand-held devices. Unlike single-image deblurring, video-based approaches can take advant age of the abundant information that exists across neighboring frames. As a resu lt the best performing methods rely on the alignment of nearby frames. However, aligning images is a computationally expensive and fragile procedure, and method s that aggregate information must therefore be able to identify which regions ha

ve been accurately aligned and which have not, a task that requires high level s
cene understanding. In this work, we introduce a deep learning solution to video
 deblurring, where a CNN is trained end-to-end to learn how to accumulate inform
ation across frames. To train this network, we collected a dataset of real video
s recorded with a high frame rate camera, which we use to generate synthetic mot
ion blur for supervision. We show that the features learned from this dataset ex
tend to deblurring motion blur that arises due to camera shake in a wide range o
f videos, and compare the quality of results to a number of other baselines.
**********

Accurate Optical Flow via Direct Cost Volume Processing
Jia Xu, Rene Ranftl, Vladlen Koltun; Proceedings of the IEEE Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2017, pp. 1289-1297
We present an optical flow estimation approach that operates on the full four-di
mensional cost volume. This direct approach shares the structural benefits of le
ading stereo matching pipelines, which are known to yield high accuracy. To this
 day, such approaches have been considered impractical due to the size of the co
st volume. We show that the full four-dimensional cost volume can be constructed
 in a fraction of a second due to its regularity. We then exploit this regularit
y further by adapting semi-global matching to the four-dimensional setting. This
 yields a pipeline that achieves significantly higher accuracy than state-of-the
-art optical flow methods while being faster than most. Our approach outperforms
 all published general-purpose optical flow methods on both Sintel and KITTI 201
5 benchmarks.
**********

Weakly Supervised Actor-Action Segmentation via Robust Multi-Task Ranking
Yan Yan, Chenliang Xu, Dawen Cai, Jason J. Corso; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1298-1307
Fine-grained activity understanding in videos has attracted considerable recent
attention with a shift from action classification to detailed actor and action u
nderstanding that provides compelling results for perceptual needs of cutting-ed
ge autonomous systems. However, current methods for detailed understanding of ac
tor and action have significant limitations: they require large amounts of finel
y labeled data, and they fail to capture any internal relationship among actors
and actions. To address these issues, in this paper, we propose a novel, robust
multi-task ranking model for weakly supervised actor-action segmentation where o
nly video-level tags are given for training samples. Our model is able to share
useful information among different actors and actions while learning a ranking m
atrix to select representative supervoxels for actors and actions respectively.
Final segmentation results are generated by a conditional random field that cons
iders various ranking scores for different video parts. Extensive experimental r
esults on the Actor-Action Dataset (A2D) demonstrate that the proposed approach
outperforms the state-of-the-art weakly supervised methods and performs as well
as the top-performing fully supervised method.
**********

Feedback Networks
Amir R. Zamir, Te-Lin Wu, Lin Sun, William B. Shen, Bertram E. Shi, Jitendra Mal
ik, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and P
attern Recognition (CVPR), 2017, pp. 1308-1317
urrently, the most successful learning models in computer vision are based on le
arning successive representations followed by a decision layer. This is usually
actualized through feedforward multilayer neural networks, e.g. ConvNets, where
each layer forms one of such successive representations. However, an alternative
 that can achieve the same goal is a feedback based approach in which the repres
entation is formed in an iterative manner based on a feedback received from prev
ious iteration's output.  We establish that a feedback based approach has severa
l core advantages over feedforward: it enables making early predictions at the q
uery time, its output naturally conforms to a hierarchical structure in the labe
l space (e.g. a taxonomy), and it provides a new basis for Curriculum Learning.
We observe that feedback develops a considerably different representation compar
ed to feedforward counterparts, in line with the aforementioned advantages. We p

rovide a general feedback based learning architecture, instantiated using existi ng RNNs, with the endpoint results on par or better than existing feedforward ne tworks and the addition of the above advantages.
**********

Re-Ranking Person Re-Identification With k-Reciprocal Encoding
Zhun Zhong, Liang Zheng, Donglin Cao, Shaozi Li; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1318-1327
When considering person re-identification (re-ID) as a retrieval process, re-ran king is a critical step to improve its accuracy. Yet in the re-ID community, lim ited effort has been devoted to re-ranking, especially those fully automatic, un supervised solutions. In this paper, we propose a k-reciprocal encoding method t o re-rank the re-ID results. Our hypothesis is that if a gallery image is simila r to the probe in the k-reciprocal nearest neighbors, it is more likely to be a true match. Specifically, given an image, a k-reciprocal feature is calculated b y encoding its k-reciprocal nearest neighbors into a single vector, which is use d for re-ranking under the Jaccard distance. The final distance is computed as t he combination of the original distance and the Jaccard distance. Our re-ranking method does not require any human interaction or any labeled data, so it is app licable to large-scale datasets. Experiments on the large-scale Market-1501, CUH K03, MARS, and PRW datasets confirm the effectiveness of our method.
**********

Deep Visual-Semantic Quantization for Efficient Image Retrieval
Yue Cao, Mingsheng Long, Jianmin Wang, Shichen Liu; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1328-1337
Compact coding has been widely applied to approximate nearest neighbor search fo r large-scale image retrieval, due to its computation efficiency and retrieval q uality. This paper presents a compact coding solution with a focus on the deep l earning to quantization approach, which improves retrieval quality by end-to-end representation learning and compact encoding and has already shown the superior performance over the hashing solutions for similarity retrieval. We propose Dee p Visual-Semantic Quantization (DVSQ), which is the first approach to learning d eep quantization models from labeled image data as well as the semantic informat ion underlying general text domains. The main contribution lies in jointly learn ing deep visual-semantic embeddings and visual-semantic quantizers using careful ly-designed hybrid networks and well-specified loss functions. DVSQ enables effi cient and effective image retrieval by supporting maximum inner-product search, which is computed based on learned codebooks with fast distance table lookup. Co mprehensive empirical evidence shows that DVSQ can generate compact binary codes and yield state-of-the-art similarity retrieval performance on standard benchma rks.
**********

Sequential Person Recognition in Photo Albums With a Recurrent Network
Yao Li, Guosheng Lin, Bohan Zhuang, Lingqiao Liu, Chunhua Shen, Anton van den He ngel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit ion (CVPR), 2017, pp. 1338-1346
Recognizing the identities of people in everyday photos is still a very challeng ing problem for machine vision, due to issues such as non-frontal faces, changes in clothing, location, lighting. Recent studies have shown that rich relational information between people in the same photo can help in recognizing their iden tities. In this work, we propose to model the relational information between peo ple as a sequence prediction task. At the core of our work is a novel recurrent network architecture, in which relational information between instances' labels and appearance are modeled jointly. In addition to relational cues, scene contex t is incorporated in our sequence prediction model with no additional cost. In t his sense, our approach is a unified framework for modeling both contextual cues and visual appearance of person instances. Our model is trained end-to-end with a sequence of annotated instances in a photo as inputs, and a sequence of corre sponding labels as targets. We demonstrate that this simple but elegant formulat ion achieves state-of-the-art performance on the newly released People In Photo Albums (PIPA) dataset.

**********
ViP-CNN: Visual Phrase Guided Convolutional Neural Network
Yikang Li, Wanli Ouyang, Xiaogang Wang, Xiao'ou Tang; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1347-1356
As the intermediate level task connecting image captioning and object detection, visual relationship detection started to catch researchers' attention because o f its descriptive power and clear structure. It detects the objects and captures their pair-wise interactions with a subject-predicate-object triplet, e.g. pers on-ride-horse. In this paper, each visual relationship is considered as a phrase with three components. We formulate the visual relationship detection as three inter-connected recognition problems and propose a Visual Phrase guided Convolut ional Neural Network (ViP-CNN) to address them simultaneously. In ViP-CNN, we pr esent a Phrase-guided Message Passing Structure (PMPS) to establish the connecti on among relationship components and help the model consider the three problems jointly. Corresponding non-maximum suppression method and model training strateg y are also proposed. Experimental results show that our ViP-CNN outperforms the state-of-art method both in speed and accuracy. We further pretrain ViP-CNN on o ur cleansed Visual Genome Relationship dataset, which is found to perform better than the pretraining on the ImageNet for this task.
**********
Deep Joint Rain Detection and Removal From a Single Image
Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, Shuicheng Yan ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1357-1366
In this paper, we address a rain removal problem from a single image, even in th e presence of heavy rain and rain streak accumulation. Our core ideas lie in our new rain image model and new deep learning architecture. We add a binary map th at provides rain streak locations to an existing model, which comprises a rain s treak layer and a background layer. We create a model consisting of a component representing rain streak accumulation (where individual streaks cannot be seen, and thus visually similar to mist or fog), and another component representing va rious shapes and directions of overlapping rain streaks, which usually happen in heavy rain. Based on the model, we develop a multi-task deep learning architect ure that learns the binary rain streak map, the appearance of rain streaks, and the clean background, which is our ultimate output. The additional binary map is critically beneficial, since its loss function can provide additional strong in formation to the network. To handle rain streak accumulation (again, a phenomeno n visually similar to mist or fog) and various shapes and directions of overlapp ing rain streaks, we propose a recurrent rain detection and removal network that removes rain streaks and clears up the rain accumulation iteratively and progre ssively. In each recurrence of our method, a new contextualized dilated network is developed to exploit regional contextual information and to produce better re presentations for rain detection. The evaluation on real images, particularly on heavy rain, shows the effectiveness of our models and architecture.
**********
Person Re-Identification in the Wild
Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ( CVPR), 2017, pp. 1367-1376
This paper presents a novel large-scale dataset and comprehensive baselines for end-to-end pedestrian detection and person recognition in raw video frames. Our baselines address three issues: the performance of various combinations of detec tors and recognizers, mechanisms for pedestrian detection to help improve overal l re-identification (re-ID) accuracy and assessing the effectiveness of differen t detectors for re-ID. We make three distinct contributions. First, a new datase t, PRW, is introduced to evaluate Person Re-identification in the Wild, using vi deos acquired through six synchronized cameras. It contains 932 identities and 1 1,816 frames in which pedestrians are annotated with their bounding box position s and identities. Extensive benchmarking results are presented on this dataset. Second, we show that pedestrian detection aids re-ID through two simple yet effe

ctive improvements: a cascaded fine-tuning strategy that trains a detection mode l first and then the classification model, and a Confidence Weighted Similarity (CWS) metric that incorporates detection scores into similarity measurement. Thi rd, we derive insights in evaluating detector performance for the particular sce nario of accurate person re-ID.

**********

Deep Self-Taught Learning for Weakly Supervised Object Localization

Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, Wei Liu; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1377-1385

Most existing weakly supervised localization (WSL) approaches learn detectors by finding positive bounding boxes based on features learned with image-level supe rvision. However, those features do not contain spatial location related informa tion and usually provide poor-quality positive samples for training a detector. To overcome this issue, we propose a deep self-taught learning approach, which m akes the detector learn the object-level features reliable for acquiring tight p ositive samples and afterwards re-train itself based on them. Consequently, the detector progressively improves its detection ability and localizes more informa tive positive samples. To implement such self-taught learning, we propose a seed sample acquisition method via image-to-object transferring and dense subgraph d iscovery to find reliable positive samples for initializing the detector. An onl ine supportive sample harvesting scheme is further proposed to dynamically selec t the most confident tight positive samples and train the detector in a mutual b oosting way. To prevent the detector from being trapped in poor optima due to ov erfitting, we propose a new relative improvement of predicted CNN scores for gui ding the self-taught learning process. Extensive experiments on PASCAL 2007 and 2012 show that our approach outperforms the state-of-the-arts, strongly validati ng its effectiveness.

**********

KillingFusion: Non-Rigid 3D Reconstruction Without Correspondences

Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, Slobodan Ilic; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 , pp. 1386-1395

We introduce a geometry-driven approach for real-time 3D reconstruction of defor ming surfaces from a single RGB-D stream without any templates or shape priors. To this end, we tackle the problem of non-rigid registration by level set evolut ion without explicit correspondence search. Given a pair of signed distance fiel ds (SDFs) representing the shapes of interest, we estimate a dense deformation f ield that aligns them. It is defined as a displacement vector field of the same resolution as the SDFs and is determined iteratively via variational minimizatio n. To ensure it generates plausible shapes, we propose a novel regularizer that imposes local rigidity by requiring the deformation to be a smooth and approxima tely Killing vector field, i.e. generating nearly isometric motions. Moreover, w e enforce that the level set property of unity gradient magnitude is preserved o ver iterations. As a result, KillingFusion reliably reconstructs objects that ar e undergoing topological changes and fast inter-frame motion. In addition to inc rementally building a model from scratch, our system can also deform complete su rfaces. We demonstrate these capabilities on several public datasets and introdu ce our own sequences that permit both qualitative and quantitative comparison to related approaches.

**********

Context-Aware Correlation Filter Tracking

Matthias Mueller, Neil Smith, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1396-1404

Correlation filter (CF) based trackers have recently gained a lot of popularity due to their impressive performance on benchmark datasets, while maintaining hig h frame rates. A significant amount of recent research focuses on the incorporat ion of stronger features for a richer representation of the tracking target. How ever, this only helps to discriminate the target from background within a small neighborhood. In this paper, we present a framework that allows the explicit inc

orporation of global context within CF trackers. We reformulate the original opt
imization problem and provide a closed form solution for single and multi-dimens
ional features in the primal and dual domain. Extensive experiments demonstrate
that this framework significantly improves the performance of many CF trackers w
ith only a modest impact on frame rate.
**********

Missing Modalities Imputation via Cascaded Residual Autoencoder
Luan Tran, Xiaoming Liu, Jiayu Zhou, Rong Jin; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1405-1414
Affordable sensors lead to an increasing interest in acquiring and modeling data
 with multiple modalities. Learning from multiple modalities has shown to signif
icantly improve performance in object recognition. However, in practice it is co
mmon that the sensing equipment experiences unforeseeable malfunction or configu
ration issues, leading to corrupted data with missing modalities. Most existing
multi-modal learning algorithms could not handle missing modalities, and would d
iscard either all modalities with missing values or all corrupted data. To lever
age the valuable information in the corrupted data, we propose to impute the mis
sing data by leveraging the relatedness among different modalities. Specifically
, we propose a novel Cascaded Residual Autoencoder (CRA) to impute missing modal
ities. By stacking residual autoencoders, CRA grows iteratively to model the res
idual between the current prediction and original data. Extensive experiments de
monstrate the superior performance of CRA on both the data imputation and the ob
ject recognition task on imputed data.
**********

Disentangled Representation Learning GAN for Pose-Invariant Face Recognition
Luan Tran, Xi Yin, Xiaoming Liu; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2017, pp. 1415-1424
The large pose discrepancy between two face images is one of the key challenges
in face recognition. Conventional approaches for pose-invariant face recognition
 either perform face frontalization on, or learn a pose-invariant representation
 from, a non-frontal face image. We argue that it is more desirable to perform b
oth tasks jointly to allow them to leverage each other. To this end, this paper
proposes Disentangled Representation learning-Generative Adversarial Network (DR
-GAN) with three distinct novelties. First, the encoder-decoder structure of the
 generator allows DR-GAN to learn a generative and discriminative representation
, in addition to image synthesis. Second, this representation is explicitly dise
ntangled from other face variations such as pose, through the pose code provided
 to the decoder and pose estimation in the discriminator. Third, DR-GAN can take
 one or multiple images as the input, and generate one unified representation al
ong with an arbitrary number of synthetic images. Quantitative and qualitative e
valuation on both controlled and in-the-wild databases demonstrate the superiori
ty of DR-GAN over the state of the art.
**********

Discretely Coding Semantic Rank Orders for Supervised Image Hashing
Li Liu, Ling Shao, Fumin Shen, Mengyang Yu; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1425-1434
Learning to hash has been recognized to accomplish highly efficient storage and
retrieval for large-scale visual data. Particularly, ranking-based hashing techn
iques have recently attracted broad research attention because ranking accuracy
among the retrieved data is well explored and their objective is more applicable
 to realistic search tasks. However, directly optimizing discrete hash codes wit
hout continuous-relaxations on a nonlinear ranking objective is infeasible by ei
ther traditional optimization methods or even recent discrete hashing algorithms
. To address this challenging issue, in this paper, we introduce a novel supervi
sed hashing method, dubbed Discrete Semantic Ranking Hashing (DSeRH), which aims
 to directly embed semantic rank orders into binary codes. In DSeRH, a generaliz
ed Adaptive Discrete Minimization (ADM) approach is proposed to discretely optim
ize binary codes with the quadratic nonlinear ranking objective in an iterative
manner and is guaranteed to converge quickly. Additionally, instead of using 0/1
 independent labels to form rank orders as in previous works, we generate the li

stwise rank orders from the high-level semantic word embeddings which can quanti tatively capture the intrinsic correlation between different categories. We eval uate our DSeRH, coupled with both linear and deep convolutional neural network ( CNN) hash functions, on three image datasets, i.e., CIFAR-10, SUN397 and ImageNe t100, and the results manifest that DSeRH can outperform the state-of-the-art ra nking-based hashing methods.
**********

NID-SLAM: Robust Monocular SLAM Using Normalised Information Distance
Geoffrey Pascoe, Will Maddern, Michael Tanner, Pedro Pinies, Paul Newman; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1435-1444
We propose a direct monocular SLAM algorithm based on the Normalised Information Distance (NID) metric. In contrast to current state-of-the-art direct methods b ased on photometric error minimisation, our information-theoretic NID metric pro vides robustness to appearance variation due to lighting, weather and structural changes in the scene. We demonstrate successful localisation and mapping across changes in lighting with a synthetic indoor scene, and across changes in weathe r (direct sun, rain, snow) using real-world data collected from a vehicle-mounte d camera. Our approach runs in real-time on a consumer GPU using OpenGL, and pro vides comparable localisation accuracy to state-of-the-art photometric methods b ut significantly outperforms both direct and feature-based methods in robustness to appearance changes.
**********

Efficient Optimization for Hierarchically-structured Interacting Segments (HINTS )
Hossam Isack, Olga Veksler, Ipek Oguz, Milan Sonka, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1445-1453
We propose an effective optimization algorithm for a general hierarchical segmen tation model with geometric interactions between segments. Any given tree can sp ecify a partial order over object labels defining a hierarchy. It is well-establ ished that segment interactions, such as inclusion/exclusion and margin constrai nts, make the model significantly more discriminant. However, existing optimizat ion methods do not allow full use of such models. Generic a-expansion results in weak local minima, while common binary multi-layered formulations lead to non-s ubmodularity, complex high-order potentials, or polar domain unwrapping and shap e biases. In practice, applying these methods to arbitrary trees does not work e xcept for simple cases. Our main contribution is an optimization method for the Hierarchically-structured Interacting Segments (HINTS) model with arbitrary tree s. Our Path-Moves algorithm is based on multi-label MRF formulation and can be s een as a combination of well-known a-expansion and Ishikawa techniques. We show state-of-the-art biomedical segmentation for many diverse examples of complex tr ees.
**********

SCC: Semantic Context Cascade for Efficient Action Detection
Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, Bernard Ghanem; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 7, pp. 1454-1463
Despite the recent advances in large-scale video analysis, action detection rema ins as one of the most challenging unsolved problems in computer vision. This sn ag is in part due to the large volume of data that needs to be analyzed to detec t actions in videos. Existing approaches have mitigated the computational cost, but still, these methods lack rich high-level semantics that helps them to local ize the actions quickly. In this paper, we introduce a Semantic Cascade Context (SCC) model that aims to detect action in long video sequences. By embracing sem antic priors associated with human activities, SCC produces high-quality class-s pecific action proposals and prune unrelated activities in a cascade fashion. Ex perimental results in ActivityNet unveils that SCC achieves state-of-the-art per formance for action detection while operating at real time.
**********

Semantic Amodal Segmentation
Yan Zhu, Yuandong Tian, Dimitris Metaxas, Piotr Dollar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1464-1472
Common visual recognition tasks such as classification, object detection, and semantic segmentation are rapidly reaching maturity, and given the recent rate of progress, it is not unreasonable to conjecture that techniques for many of these problems will approach human levels of performance in the next few years. In this paper we look to the future: what is the next frontier in visual recognition? We offer one possible answer to this question. We propose a detailed image annotation that captures information beyond the visible pixels and requires complex reasoning about full scene structure. Specifically, we create an amodal segmentation of each image: the full extent of each region is marked, not just the visible pixels. Annotators outline and name all salient regions in the image and specify a partial depth order. The result is a rich scene structure, including visible and occluded portions of each region, figure-ground edge information, semantic labels, and object overlap. We create two datasets for semantic amodal segmentation. First, we label 500 images in the BSDS dataset with multiple annotators per image, allowing us to study the statistics of human annotations. We show that the proposed full scene annotation is surprisingly consistent between annotators, including for regions and edges. Second, we annotate 5000 images from COCO. This larger dataset allows us to explore a number of algorithmic ideas for amodal segmentation and depth ordering. We introduce novel metrics for these tasks, and along with our strong baselines, define concrete new challenges for the community.
**********
Deep Sequential Context Networks for Action Prediction
Yu Kong, Zhiqiang Tao, Yun Fu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1473-1481
This paper proposes efficient and powerful deep networks for action prediction from partially observed videos containing temporally incomplete action executions. Different from after-the-fact action recognition, action prediction task requires action labels to be predicted from these partially observed videos. Our approach exploits abundant sequential context information to enrich the feature representations of partial videos. We reconstruct missing information in the features extracted from partial videos by learning from fully observed action videos. The amount of the information is temporally ordered for the purpose of modeling temporal orderings of action segments. Label information is also used to better separate the learned features of different categories. We develop a new learning formulation that enables efficient model training. Extensive experimental results on UCF101, Sports-1M and BIT datasets demonstrate that our approach remarkably outperforms state-of-the-art methods, and is up to 300x faster than these methods. Results also show that actions differ in their prediction characteristics; some actions can be correctly predicted even though only the beginning 10% portion of videos is observed.
**********
Comparative Evaluation of Hand-Crafted and Learned Local Features
Johannes L. Schonberger, Hans Hardmeier, Torsten Sattler, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1482-1491
Matching local image descriptors is a key step in many computer vision applications. For more than a decade, hand-crafted descriptors such as SIFT have been used for this task. Recently, multiple new descriptors learned from data have been proposed and shown to improve on SIFT in terms of discriminative power. This paper is dedicated to an extensive experimental evaluation of learned local features to establish a single evaluation protocol that ensures comparable results. In terms of matching performance, we evaluate the different descriptors regarding standard criteria. However, considering matching performance in isolation only provides an incomplete measure of a descriptor's quality. For example, finding additional correct matches between similar images does not necessarily lead to a be

tter performance when trying to match images under extreme viewpoint or illumination changes. Besides pure descriptor matching, we thus also evaluate the different descriptors in the context of image-based reconstruction. This enables us to study the descriptor performance on a set of more practical criteria including image retrieval, the ability to register images under strong viewpoint and illumination changes, and the accuracy and completeness of the reconstructed cameras and scenes. To facilitate future research, the full evaluation pipeline is made publicly available.
**********

Aggregated Residual Transformations for Deep Neural Networks
Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, Kaiming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1492-1500
We present a simple, highly modularized network architecture for image classification. Our network is constructed by repeating a building block that aggregates a set of transformations with the same topology. Our simple design results in a homogeneous, multi-branch architecture that has only a few hyper-parameters to set. This strategy exposes a new dimension, which we call "cardinality" (the size of the set of transformations), as an essential factor in addition to the dimensions of depth and width. On the ImageNet-1K dataset, we empirically show that even under the restricted condition of maintaining complexity, increasing cardinality is able to improve classification accuracy. Moreover, increasing cardinality is more effective than going deeper or wider when we increase the capacity. Our models, named ResNeXt, are the foundations of our entry to the ILSVRC 2016 classification task in which we secured 2nd place. We further investigate ResNeXt on an ImageNet-5K set and the COCO detection set, also showing better results than its ResNet counterpart. The code and models are publicly available online.
**********

Predicting Behaviors of Basketball Players From First Person Videos
Shan Su, Jung Pyo Hong, Jianbo Shi, Hyun Soo Park; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1501-1510
This paper presents a method to predict the future movements (location and gaze direction) of basketball players as a whole from their first person videos. The predicted behaviors reflect an individual physical space that affords to take the next actions while conforming to social behaviors by engaging to joint attention. Our key innovation is to use the 3D reconstruction of multiple first person cameras to automatically annotate each other's visual semantics of social configurations. We leverage two learning signals uniquely embedded in first person videos. Individually, a first person video records the visual semantics of a spatial and social layout around a person that allows associating with past similar situations. Collectively, first person videos follow joint attention that can link the individuals to a group. We learn the egocentric visual semantics of group movements using a Siamese neural network to retrieve future trajectories. We consolidate the retrieved trajectories from all players by maximizing a measure of social compatibility---the gaze alignment towards joint attention predicted by their social formation, where the dynamics of joint attention is learned by a long-term recurrent convolutional network. This allows us to characterize which social configuration is more plausible and predict future group trajectories.
**********

Synthesizing 3D Shapes via Modeling Multi-View Depth Maps and Silhouettes With Deep Generative Networks
Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D. Kulkarni, Joshua B. Tenenbaum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1511-1519
We study the problem of learning generative models of 3D shapes. Voxels or 3D parts have been widely used as the underlying representations to build complex 3D shapes; however, voxel-based representations suffer from high memory requirements, and parts-based models require a large collection of cached or richly parametrized parts. We take an alternative approach: learning a generative model over multi-view depth maps or their corresponding silhouettes, and using a determinist

ic rendering function to produce 3D shapes from these images. A multi-view repre
sentation of shapes enables generation of 3D models with fine details, as 2D dep
th maps and silhouettes can be modeled at a much higher resolution than 3D voxel
s. Moreover, our approach naturally brings the ability to recover the underlying
 3D representation from depth maps of one or a few viewpoints. Experiments show
that our framework can generate 3D shapes with variations and details. We also d
emonstrate that our model has out-of-sample generalization power for real-world
tasks with occluded objects.
**********

Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search
Bo Zhao, Jiashi Feng, Xiao Wu, Shuicheng Yan; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1520-1528
We introduce a new fashion search protocol where attribute manipulation is allow
ed within the interaction between users and search engines, e.g. manipulating th
e color attribute of the clothing from red to blue. It is particularly useful fo
r image-based search when the query image cannot perfectly match user's expectat
ion of the desired product. To build such a search engine, we propose a novel me
mory-augmented Attribute Manipulation Network (AMNet) which can manipulate image
 representation at the attribute level. Given a query image and some attributes
that need to modify, AMNet can manipulate the intermediate representation encodi
ng the unwanted attributes and change them to the desired ones through following
 four novel components: (1) a dual-path CNN architecture for discriminative deep
 attribute representation learning; (2) a memory block with an internal memory a
nd a neural controller for prototype attribute representation learning and hosti
ng; (3) an attribute manipulation network to modify the representation of the qu
ery image with the prototype feature retrieved from the memory block; (4) a loss
 layer which jointly optimizes the attribute classification loss and a triplet r
anking loss over triplet images for facilitating precise attribute manipulation
and image retrieving. Extensive experiments conducted on two large-scale fashion
 search datasets, i.e. DARN and DeepFashion, have demonstrated that AMNet is abl
e to achieve remarkably good performance compared with well-designed baselines i
n terms of effectiveness of attribute manipulation and search accuracy.
**********

Spatiotemporal Pyramid Network for Video Action Recognition
Yunbo Wang, Mingsheng Long, Jianmin Wang, Philip S. Yu; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1529-153
8
Two-stream convolutional networks have shown strong performance in video action
recognition tasks. The key idea is to learn spatiotemporal features by fusing co
nvolutional networks spatially and temporally. However, it remains unclear how t
o model the correlations between the spatial and temporal structures at multiple
 abstraction levels. First, the spatial stream tends to fail if two videos share
 similar backgrounds. Second, the temporal stream may be fooled if two actions r
esemble in short snippets, though appear to be distinct in the long term. We pro
pose a novel spatiotemporal pyramid network to fuse the spatial and temporal fea
tures in a pyramid structure such that they can reinforce each other. From the a
rchitecture perspective, our network constitutes hierarchical fusion strategies
which can be trained as a whole using a unified spatiotemporal loss. A series of
 ablation experiments support the importance of each fusion strategy. From the t
echnical perspective, we introduce the spatiotemporal compact bilinear operator
into video analysis tasks. This operator enables efficient training of bilinear
fusion operations which can capture full interactions between the spatial and te
mporal features. Our final network achieves state-of-the-art results on standard
 video datasets.
**********

Reconstructing Transient Images From Single-Photon Sensors
Matthew O'Toole, Felix Heide, David B. Lindell, Kai Zang, Steven Diamond, Gordon
 Wetzstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Re
cognition (CVPR), 2017, pp. 1539-1547
Computer vision algorithms build on 2D images or 3D videos that capture dynamic

events at the millisecond time scale. However, capturing and analyzing "transien t images" at the picosecond scale---i.e., at one trillion frames per second---re veals unprecedented information about a scene and light transport within. This i s not only crucial for time-of-flight range imaging, but it also helps further o ur understanding of light transport phenomena at a more fundamental level and po tentially allows to revisit many assumptions made in different computer vision a lgorithms.  In this work, we design and evaluate an imaging system that builds o n single photon avalanche diode (SPAD) sensors to capture multi-path responses w ith picosecond-scale active illumination. We develop inverse methods that use mo dern approaches to deconvolve and denoise measurements in the presence of Poisso n noise, and compute transient images at a higher quality than previously report ed. The small form factor, fast acquisition rates, and relatively low cost of ou r system potentially makes transient imaging more practical for a range of appli cations.

**********

Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network

Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1548-1557

Facial analysis in videos, including head pose estimation and facial landmark lo calization, is key for many applications such as facial animation capture, human activity recognition, and human-computer interaction. In this paper, we propose to use a recurrent neural network (RNN) for joint estimation and tracking of fa cial features in videos. We are inspired by the fact that the computation perfor med in an RNN bears resemblance to Bayesian filters, which have been used for tr acking in many previous methods for facial analysis from videos. Bayesian filter s used in these methods, however, require complicated, problem-specific design a nd tuning. In contrast, our proposed RNN-based method avoids such tracker-engine ering by learning from training data, similar to how a convolutional neural netw ork (CNN) avoids feature-engineering for image classification. As an end-to-end network, the proposed RNN-based method provides a generic and holistic solution for joint estimation and tracking of various types of facial features from conse cutive video frames. Extensive experimental results on head pose estimation and facial landmark localization from videos demonstrate that the proposed RNN-based method outperforms frame-wise models and Bayesian filtering. In addition, we cr eate a large-scale synthetic dataset for head pose estimation, with which we ach ieve state-of-the-art performance on a benchmark dataset.

**********

Polarimetric Multi-View Stereo

Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1558-15 67

Multi-view stereo relies on feature correspondences for 3D reconstruction, and t hus is fundamentally flawed in dealing with featureless scenes. In this paper, w e propose polarimetric multi-view stereo, which combines per-pixel photometric i nformation from polarization with epipolar constraints from multiple views for 3 D reconstruction. Polarization reveals surface normal information, and is thus h elpful to propagate depth to featureless regions. Polarimetric multi-view stereo is completely passive and can be applied outdoors in uncontrolled illumination, since the data capture can be done simply with either a polarizer or a polariza tion camera. Unlike previous work on shape-from-polarization which is limited to either diffuse polarization or specular polarization only, we propose a novel p olarization imaging model that can handle real-world objects with mixed polariza tion. We prove there are exactly two types of ambiguities on estimating surface azimuth angles from polarization, and we resolve them with graph optimization an d iso-depth contour tracing. This step significantly improves the initial depth map estimate, which are later fused together for complete 3D reconstruction. Ext ensive experimental results demonstrate high-quality 3D reconstruction and bette r performance than state-of-the-art multi-view stereo methods, especially on fea tureless 3D objects, such as ceramic tiles, office room with white walls, and hi ghly reflective cars in the outdoors.

**********
Object Region Mining With Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach
Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1568-1576
We investigate a principle way to progressively mine discriminative object regions using classification networks to address the weakly-supervised semantic segmentation problems. Classification networks are only responsive to small and sparse discriminative regions from the object of interest, which deviates from the requirement of the segmentation task that needs to localize dense, interior and integral regions for pixel-wise inference. To mitigate this gap, we propose a new adversarial erasing approach for localizing and expanding object regions progressively. Starting with a single small object region, our proposed approach drives the classification network to sequentially discover new and complement object regions by erasing the current mined regions in an adversarial manner. These localized regions eventually constitute a dense and complete object region for learning semantic segmentation. To further enhance the quality of the discovered regions by adversarial erasing, an online prohibitive segmentation learning approach is developed to collaborate with adversarial erasing by providing auxiliary segmentation supervision modulated by the more reliable classification scores. Despite its apparent simplicity, the proposed approach achieves 55.0% and 55.7% mean Intersection-over-Union (mIoU) scores on PASCAL VOC 2012 val and test sets, which are the new state-of-the-arts.
**********
MIML-FCN+: Multi-Instance Multi-Label Learning via Fully Convolutional Networks With Privileged Information
Hao Yang, Joey Tianyi Zhou, Jianfei Cai, Yew Soon Ong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1577-1585
Multi-instance multi-label (MIML) learning has many interesting applications in computer visions, including multi-object recognition and automatic image tagging. In these applications, additional information such as bounding-boxes, image captions and descriptions is often available during training phrase, which is referred as privileged information (PI). However, as existing works on learning using PI only consider instance-level PI (privileged instances), they fail to make use of bag-level PI (privileged bags) available in MIML learning. Therefore, in this paper, we propose a two-stream fully convolutional network, named MIML-FCN+, unified by a novel PI loss to solve the problem of MIML learning with privileged bags. Compared to the previous works on PI, the proposed MIML-FCN+ utilizes the readily available privileged bags, instead of hard-to-obtain privileged instances, making the system more general and practical in real world applications. As the proposed PI loss is convex and SGD-compatible and the framework itself is a fully convolutional network, MIML FCN+ can be easily integrated with state-of-the-art deep learning networks. Moreover, the flexibility of convolutional layers allows us to exploit structured correlations among instances to facilitate more effective training and testing. Experimental results on three benchmark datasets demonstrate the effectiveness of the proposed MIML-FCN+, outperforming state-of-the-art methods in the application of multi-object recognition.
**********
Benchmarking Denoising Algorithms With Real Photographs
Tobias Plotz, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1586-1595
Lacking realistic ground truth data, image denoising techniques are traditionally evaluated on images corrupted by synthesized i.i.d. Gaussian noise. We aim to obviate this unrealistic setting by developing a methodology for benchmarking denoising techniques on real photographs. We capture pairs of images with different ISO values and appropriately adjusted exposure times, where the nearly noise-free low-ISO image serves as reference. To derive the ground truth, careful post-processing is needed. We correct spatial misalignment, cope with inaccuracies in the exposure parameters through a linear intensity transform based on a novel h

eteroscedastic Tobit regression model, and remove residual low-frequency bias th
at stems, e.g., from minor illumination changes. We then capture a novel benchma
rk dataset, the Darmstadt Noise Dataset (DND), with consumer cameras of differin
g sensor sizes. One interesting finding is that various recent techniques that p
erform well on synthetic noise are clearly outperformed by BM3D on photographs w
ith real noise. Our benchmark delineates realistic evaluation scenarios that dev
iate strongly from those commonly used in the scientific literature.
**********

A Dual Ascent Framework for Lagrangean Decomposition of Combinatorial Problems
Paul Swoboda, Jan Kuske, Bogdan Savchynskyy; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1596-1606
We propose a general dual ascent (message passing) framework for Lagrangean (dua
l) decomposition of combinatorial problems. Although methods of this type have s
hown their efficiency for a number of problems, so far there was no general algo
rithm applicable to multiple problem types. In this work, we propose such a gene
ral algorithm. It depends on several parameters, which can be used to optimize i
ts performance in each particular setting. We demonstrate efficiency of our meth
od on the graph matching and the multicut problems, where it outperforms state-o
f-the-art solvers including those based on the subgradient optimization and off-
the-shelf linear programming solvers.
**********

A Study of Lagrangean Decompositions and Dual Ascent Solvers for Graph Matching
Paul Swoboda, Carsten Rother, Hassan Abu Alhaija, Dagmar Kainmuller, Bogdan Savc
hynskyy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recog
nition (CVPR), 2017, pp. 1607-1616
We study the quadratic assignment problem, in computer vision also known as grap
h matching. Two leading solvers for this problem optimize the Lagrange decomposi
tion duals with sub-gradient and dual ascent (also known as message passing) upd
ates. We explore this direction further and propose several additional Lagrangea
n relaxations of the graph matching problem along with corresponding algorithms,
 which are all based on a common dual ascent framework. Our extensive empirical
evaluation gives several theoretical insights and suggests a new state-of-the-ar
t anytime solver for the considered problem. Our improvement over state-of-the-a
rt is particularly visible on a new dataset with large-scale sparse problem inst
ances containing more than 500 graph nodes each.
**********

A Message Passing Algorithm for the Minimum Cost Multicut Problem
Paul Swoboda, Bjoern Andres; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2017, pp. 1617-1626
We propose a dual decomposition and linear program relaxation of the NP-hard min
imum cost multicut problem. Unlike other polyhedral relaxations of the multicut
polytope, it is amenable to efficient optimization by message passing. Like othe
r polyhedral relaxations, it can be tightened efficiently by cutting planes. We
define an algorithm that alternates between message passing and efficient separa
tion of cycle- and odd-wheel inequalities. This algorithm is more efficient than
 state-of-the-art algorithms based on linear programming, including algorithms w
ritten in the framework of leading commercial software, as we show in experiment
s with large instances of the problem from applications in computer vision, biom
edical image analysis and data mining.
**********

From Zero-Shot Learning to Conventional Supervised Classification: Unseen Visual
 Data Synthesis
Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, Jungong Han; Proceeding
s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
, pp. 1627-1636
Robust object recognition systems usually rely on powerful feature extraction me
chanisms from a large number of real images. However, in many realistic applicat
ions, collecting sufficient images for ever-growing new classes is unattainable.
 In this paper, we propose a new Zero-shot learning (ZSL) framework that can syn
thesise visual features for unseen classes without acquiring real images. Using

the proposed Unseen Visual Data Synthesis (UVDS) algorithm, semantic attributes are effectively utilised as an intermediate clue to synthesise unseen visual features at the training stage. Hereafter, ZSL recognition is converted into the conventional supervised problem, i.e. the synthesised visual features can be straightforwardly fed to typical classifiers such as SVM. On four benchmark datasets, we demonstrate the benefit of using synthesised unseen data. Extensive experimental results manifest that our proposed approach significantly improve the state-of-the-art results.
**********

Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?
Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1637-1646
Accurate visual localization is a key technology for autonomous navigation. 3D structure-based methods employ 3D models of the scene to estimate the full 6DOF pose of a camera very accurately. However, constructing (and extending) large-scale 3D models is still a significant challenge. In contrast, 2D image retrieval-based methods only require a database of geo-tagged images, which is trivial to construct and to maintain. They are often considered inaccurate since they only approximate the positions of the cameras. Yet, the exact camera pose can theoretically be recovered when enough relevant database images are retrieved. In this paper, we demonstrate experimentally that large-scale 3D models are not strictly necessary for accurate visual localization. We create reference poses for a large and challenging urban dataset. Using these poses, we show that combining image-based methods with local reconstructions results in a pose accuracy similar to the state-of-the-art structure-based methods. Our results suggest that we might want to reconsider the current approach for accurate large-scale localization.
**********

Global Context-Aware Attention LSTM Networks for 3D Action Recognition
Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, Alex C. Kot; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1647-1656
Long Short-Term Memory (LSTM) networks have shown superior performance in 3D human action recognition due to their power in modeling the dynamics and dependencies in sequential data. Since not all joints are informative for action analysis and the irrelevant joints often bring a lot of noise, we need to pay more attention to the informative ones. However, original LSTM does not have strong attention capability. Hence we propose a new class of LSTM network, Global Context-Aware Attention LSTM (GCA-LSTM), for 3D action recognition, which is able to selectively focus on the informative joints in the action sequence with the assistance of global contextual information. In order to achieve a reliable attention representation for the action sequence, we further propose a recurrent attention mechanism for our GCA-LSTM network, in which the attention performance is improved iteratively. Experiments show that our end-to-end network can reliably focus on the most informative joints in each frame of the skeleton sequence. Moreover, our network yields state-of-the-art performance on three challenging datasets for 3D action recognition.
**********

Hierarchical Boundary-Aware Neural Encoder for Video Captioning
Lorenzo Baraldi, Costantino Grana, Rita Cucchiara; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1657-1666
The use of Recurrent Neural Networks for video captioning has recently gained a lot of attention, since they can be used both to encode the input video and to generate the corresponding description. In this paper, we present a recurrent video encoding scheme which can discover and leverage the hierarchical structure of the video. Unlike the classical encoder-decoder approach, in which a video is encoded continuously by a recurrent layer, we propose a novel LSTM cell which can identify discontinuity points between frames or segments and modify the temporal connections of the encoding layer accordingly. We evaluate our approach on three large-scale datasets: the Montreal Video Annotation dataset, the MPII Movie D

escription dataset and the Microsoft Video Description Corpus. Experiments show that our approach can discover appropriate hierarchical representations of input videos and improve the state of the art results on movie description datasets.
**********

Emotion Recognition in Context
Ronak Kosti, Jose M. Alvarez, Adria Recasens, Agata Lapedriza; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1667-1675
Understanding what a person is experiencing from her frame of reference is essential in our everyday life. For this reason, one can think that machines with this type of ability would interact better with people. However, there are no current systems capable of understanding in detail people's emotional states. Previous research on computer vision to recognize emotions has mainly focused on analyzing the facial expression, usually classifying it into the 6 basic emotions [11]. However, the context plays an important role in emotion perception, and when the context is incorporated, we can infer more emotional states. In this paper we present the Emotions in Context Database (EMCO), a dataset of images containing people in context in non-controlled environments. In these images, people are annotated with 26 emotional categories and also with the continuous dimensions valence, arousal, and dominance [21]. With the EMCO dataset, we trained a Convolutional Neural Network model that jointly analyzes the person and the whole scene to recognize rich information about emotional states. With this, we show the importance of considering the context for recognizing people's emotions in images, and provide a benchmark in the task of emotion recognition in visual context.
**********

Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework
Jongyoo Kim, Sanghoon Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1676-1684
Since human observers are the ultimate receivers of digital images, image quality metrics should be designed from a human-oriented perspective. Conventionally, a number of full-reference image quality assessment (FR-IQA) methods adopted various computational models of the human visual system (HVS) from psychological vision science research. In this paper, we propose a novel convolutional neural networks (CNN) based FR-IQA model, named Deep Image Quality Assessment (DeepQA), where the behavior of the HVS is learned from the underlying data distribution of IQA databases. Different from previous studies, our model seeks the optimal visual weight based on understanding of database information itself without any prior knowledge of the HVS. Through the experiments, we show that the predicted visual sensitivity maps agree with the human subjective opinions. In addition, DeepQA achieves the state-of-the-art prediction accuracy among FR-IQA models.
**********

Learning Non-Lambertian Object Intrinsics Across ShapeNet Categories
Jian Shi, Yue Dong, Hao Su, Stella X. Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1685-1694
We focus on the non-Lambertian object-level intrinsic problem of recovering diffuse albedo, shading, and specular highlights from a single image of an object. Based on existing 3D models in the ShapeNet database, a large-scale object intrinsics database is rendered with HDR environment maps. Millions of synthetic images of objects and their corresponding albedo, shading, and specular ground-truth images are used to train an encoder-decoder CNN, which can decompose an image into the product of albedo and shading components along with an additive specular component. Our CNN delivers accurate and sharp results in this classical inverse problem of computer vision. Evaluated on our realistically synthetic dataset, our method consistently outperforms the state-of-the-art by a large margin. We train and test our CNN across different object categories. Perhaps surprising especially from the CNN classification perspective, our intrinsics CNN generalizes very well across categories. Our analysis shows that feature learning at the encoder stage is more crucial for developing a universal representation across categories. We apply our model to real images and videos from Internet, and observe robust and realistic intrinsics results. Quality non-Lambertian intrinsics could

open up many interesting applications such as realistic product search based on material properties and image-based albedo / specular editing.
**********

Collaborative Deep Reinforcement Learning for Joint Object Search
Xiangyu Kong, Bo Xin, Yizhou Wang, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1695-1704
We examine the problem of joint top-down active search of multiple objects under interaction, e.g., person riding a bicycle, cups held by the table, etc.. Such objects under interaction often can provide contextual cues to each other to facilitate more efficient search. By treating each detector as an agent, we present the first collaborative multi-agent deep reinforcement learning algorithm to learn the optimal policy for joint active object localization, which effectively exploits such beneficial contextual information. We learn inter-agent communication through cross connections with gates between the Q-networks, which is facilitated by a novel multi-agent deep Q-learning algorithm with joint exploitation sampling. We verify our proposed method on multiple object detection benchmarks. Not only does our model help to improve the performance of state-of-the-art active localization models, it also reveals interesting co-detection patterns that are intuitively interpretable.
**********

Automatic Understanding of Image and Video Advertisements
Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, Adriana Kovashka; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1705-1715
There is more to images than their objective physical content: for example, advertisements are created to persuade a viewer to take a certain action. We propose the novel problem of automatic advertisement understanding. To enable research on this problem, we create two datasets: an image dataset of 64,832 image ads, and a video dataset of 3,477 ads. Our data contains rich annotations encompassing the topic and sentiment of the ads, questions and answers describing what actions the viewer is prompted to take and the reasoning that the ad presents to persuade the viewer ("What should I do according to this ad, and why should I do it?"), and symbolic references ads make (e.g. a dove symbolizes peace). We also analyze the most common persuasive strategies ads use, and the capabilities that computer vision systems should have to understand these strategies. We present baseline classification results for several prediction tasks, including automatically answering questions about the messages of the ads.
**********

FFTLasso: Large-Scale LASSO in the Fourier Domain
Adel Bibi, Hani Itani, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1716-1725
In this paper, we revisit the LASSO sparse representation problem, which has been studied and used in a variety of different areas, ranging from signal processing and information theory to computer vision and machine learning. In the vision community, it found its way into many important applications, including face recognition, tracking, super resolution, image denoising, to name a few. Despite advances in efficient sparse algorithms, solving large-scale LASSO problems remains a challenge. To circumvent this difficulty, people tend to downsample and subsample the problem (e.g. via dimensionality reduction) to maintain a manageable sized LASSO, which usually comes at the cost of losing solution accuracy. This paper proposes a novel circulant reformulation of the LASSO that lifts the problem to a higher dimension, where ADMM can be efficiently applied to its dual form. Because of this lifting, all optimization variables are updated using only basic element-wise operations, the most computationally expensive of which is a 1D FFT. In this way, there is no need for a linear system solver nor matrix-vector multiplication. Since all operations in our FFTLasso method are element-wise, the subproblems are completely independent and can be trivially parallelized (e.g. on a GPU). The attractive computational properties of FFTLasso are verified by extensive experiments on synthetic and real data and on the face recognition task. They demonstrate that FFTLasso scales much more effectively than a state-of-th

e-art solver.
**********
Multi-Modal Mean-Fields via Cardinality-Based Clamping
Pierre Baque, Francois Fleuret, Pascal Fua; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1726-1735
Mean Field inference is central to statistical physics. It has attracted much i nterest in the Computer Vision community to efficiently solve problems expressi ble in terms of large Conditional Random Fields. However, since it models the p osterior probability distribution as a product of marginal probabilities, it ma y fail to properly account for important dependencies between variables. We the refore replace the fully factorized distribution of Mean Field by a weighted mi xture of such distributions, that similarly minimizes the KL-Divergence to the true posterior. By introducing two new ideas, namely, conditioning on groups o f variables instead of single ones and using a parameter of the conditional ran dom field potentials, that we identify to the temperature in the sense of stati stical physics to select such groups, we can perform this minimization efficien tly. Our extension of the clamping method proposed in previous works allows us to both produce a more descriptive approximation of the true posterior and, ins pired by the diverse MAP paradigms, fit a mixture of Mean Field approximations. We demonstrate that this positively impacts real-world algorithms that initial ly relied on mean fields.
**********
A Unified Approach of Multi-Scale Deep and Hand-Crafted Features for Defocus Est imation
Jinsun Park, Yu-Wing Tai, Donghyeon Cho, In So Kweon; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1736-1745
In this paper, we introduce robust and synergetic hand-crafted features and a si mple but efficient deep feature from a convolutional neural network (CNN) archit ecture for defocus estimation. This paper systematically analyzes the effectiven ess of different features, and shows how each feature can compensate for the wea knesses of other features when they are concatenated. For a full defocus map est imation, we extract image patches on strong edges sparsely, after which we use t hem for deep and hand-crafted feature extraction. In order to reduce the degree of patch-scale dependency, we also propose a multi-scale patch extraction strate gy. A sparse defocus map is generated using a neural network classifier followed by a probability-joint bilateral filter. The final defocus map is obtained from the sparse defocus map with guidance from an edge-preserving filtered input ima ge. Experimental results show that our algorithm is superior to state-of-the-art algorithms in terms of defocus estimation. Our work can be used for application s such as segmentation, blur magnification, all-in-focus image generation, and 3 -D estimation.
**********
Semantic Scene Completion From a Single Depth Image
Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, Thomas Funkhou ser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogniti on (CVPR), 2017, pp. 1746-1754
This paper focuses on semantic scene completion, a task for producing a complete 3D voxel representation of volumetric occupancy and semantic labels for a scene from a single-view depth map observation. Previous work has considered scene co mpletion and semantic labeling of depth maps separately. However, we observe tha t these two problems are tightly intertwined. To leverage the coupled nature of these two tasks, we introduce the semantic scene completion network (SSCNet), an end-to-end 3D convolutional network that takes a single depth image as input an d simultaneously outputs occupancy and semantic labels for all voxels in the cam era view frustum. Our network uses a dilation-based 3D context module to efficie ntly expand the receptive field and enable 3D context learning. To train our net work, we construct SUNCG - a manually created largescale dataset of synthetic 3D scenes with dense volumetric annotations. Our experiments demonstrate that the joint model outperforms methods addressing each task in isolation and outperform s alternative approaches on the semantic scene completion task. The dataset and

code is available at http://sscnet.cs.princeton.edu.
**********
Fine-To-Coarse Global Registration of RGB-D Scans
Maciej Halber, Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1755-1764
RGB-D scanning of indoor environments is important for many applications, including real estate, interior design, and virtual reality. However, it is still challenging to register RGB-D images from a hand-held camera over a long video sequence into a globally consistent 3D model. Current methods often can lose tracking or drift and thus fail to reconstruct salient structures in large environments (e.g., parallel walls in different rooms). To address this problem, we propose a "fine-to-coarse" global registration algorithm that leverages robust registrations at finer scales to seed detection and enforcement of new correspondence and structural constraints at coarser scales. To test global registration algorithms, we provide a benchmark with 10,401 manually-clicked point correspondences in 25 scenes from the SUN3D dataset. During experiments with this benchmark, we find that our fine-to-coarse algorithm registers long RGB-D sequences better than previous methods.
**********
Universal Adversarial Perturbations
Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1765-1773
Given a state-of-the-art deep neural network classifier, we show the existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability. We propose a systematic algorithm for computing universal perturbations, and show that state-of-the-art deep neural networks are highly vulnerable to such perturbations, albeit being quasi-imperceptible to the human eye. We further empirically analyze these universal perturbations and show, in particular, that they generalize very well across neural networks. The surprising existence of universal perturbations reveals important geometric correlations among the high-dimensional decision boundary of classifiers. It further outlines potential security breaches with the existence of single directions in the input space that adversaries can possibly exploit to break a classifier on most natural images.
**********
Saliency Revisited: Analysis of Mouse Movements Versus Fixations
Hamed R. Tavakoli, Fawad Ahmed, Ali Borji, Jorma Laaksonen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1774-1782
This paper revisits visual saliency prediction by evaluating the recent advancements in this field such as crowd-sourced mouse tracking-based databases and contextual annotations. We pursue a critical and quantitative approach towards some of the new challenges including the quality of mouse tracking versus eye tracking for model training and evaluation. We extend quantitative evaluation of models in order to incorporate contextual information by proposing an evaluation methodology that allows accounting for contextual factors such as text, faces, and object attributes. The proposed contextual evaluation scheme facilitates detailed analysis of models and helps identify their pros and cons. Through several experiments, we find that (1) mouse tracking data has lower inter-participant visual congruency and higher dispersion, compared to the eye tracking data, (2) mouse tracking data does not totally agree with eye tracking in general and in terms of different contextual regions in specific, and (3) mouse tracking data leads to acceptable results in training current existing models, and (4) mouse tracking data is less reliable for model selection and evaluation. The contextual evaluation also reveals that, among the studied models, there is no single model that performs best on all the tested annotations.
**********
Online Summarization via Submodular and Convex Optimization
Ehsan Elhamifar, M. Clara De Paolis Kaluza; Proceedings of the IEEE Conference o

n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1783-1791
We consider the problem of subset selection in the online setting, where data ar rive incrementally. Instead of storing and running subset selection on the entir e dataset, we propose an incremental subset selection framework that, at each ti me instant, uses the previously selected set of representatives and the new batc h of data in order to update the set of representatives. We cast the problem as an integer binary optimization minimizing the encoding cost of the data via repr esentatives regularized by the number of selected items. As the proposed optimiz ation is, in general, NP-hard and non-convex, we study a greedy approach based o n unconstrained submodular optimization and also propose an efficient convex rel axation. We show that, under appropriate conditions, the solution of our propose d convex algorithm achieves the global optimal solution of the non-convex proble m. Our results also address the conventional problem of subset selection in the offline setting, as a special case. By extensive experiments on the problem of v ideo summarization, we demonstrate that our proposed online subset selection alg orithms perform well on real data, capturing diverse representative events in vi deos, while they obtain objective function values close to the offline setting.
**********

From Red Wine to Red Tomato: Composition With Context
Ishan Misra, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1792-1801
Compositionality and contextuality are key building blocks of intelligence. They allow us to compose known concepts to generate new and complex ones. However, t raditional learning methods do not model both these properties and require copio us amounts of labeled data to learn new concepts. A large fraction of existing t echniques, e.g., using late fusion, compose concepts but fail to model contextua lity. For example, red in red wine is different from red in red tomatoes. In thi s paper, we present a simple method that respects contextuality in order to comp ose classifiers of known visual concepts. Our method builds upon the intuition t hat classifiers lie in a smooth space where compositional transforms can be mode led. We show how it can generalize to unseen combinations of concepts. Our resul ts on composing attributes, objects as well as composing subject, predicate, and objects demonstrate its strong generalization performance compared to baselines . Finally, we present detailed analysis of our method and highlight its properti es.
**********

3DMatch: Learning Local Geometric Descriptors From RGB-D Reconstructions
Andy Zeng, Shuran Song, Matthias Niessner, Matthew Fisher, Jianxiong Xiao, Thoma s Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1802-1811
Matching local geometric features on real-world depth images is a challenging ta sk due to the noisy, low-resolution, and incomplete nature of 3D scan data. Thes e difficulties limit the performance of current state-of-art methods, which are typically based on histograms over geometric properties. In this paper, we prese nt 3DMatch, a data-driven model that learns a local volumetric patch descriptor for establishing correspondences between partial 3D data. To amass training data for our model, we propose a self-supervised feature learning method that levera ges the millions of correspondence labels found in existing RGB-D reconstruction s. Experiments show that our descriptor is not only able to match local geometry in new scenes for reconstruction, but also generalize to different tasks and sp atial scales (e.g. instance-level object model alignment for the Amazon Picking Challenge, and mesh surface correspondence). Results show that 3DMatch consisten tly outperforms other state-of-the-art approaches by a significant margin. Code, data, benchmarks, and pre-trained models are available online at http://3dmatch .cs.princeton.edu
**********

Superpixel-Based Tracking-By-Segmentation Using Markov Chains
Donghun Yeo, Jeany Son, Bohyung Han, Joon Hee Han; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1812-1821
We propose a simple but effective tracking-by-segmentation algorithm using Absor

bing Markov Chain (AMC) on superpixel segmentation, where target state is estima ted by a combination of bottom-up and top-down approaches, and target segmentati on is propagated to subsequent frames in a recursive manner. Our algorithm const ructs a graph for AMC using the superpixels identified in two consecutive frames , where background superpixels in the previous frame correspond to absorbing ver tices while all other superpixels create transient ones. The weight of each edge depends on the similarity of scores in the end superpixels, which are learned b y support vector regression. Once graph construction is completed, target segmen tation is estimated using the absorption time of each superpixel. The proposed t racking algorithm achieves substantially improved performance compared to the st ate-of-the-art segmentation-based tracking techniques in multiple challenging da tasets.

**********

Quad-Networks: Unsupervised Learning to Rank for Interest Point Detection

Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, Marc Pollefeys; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 1822-1830

Several machine learning tasks require to represent the data using only a sparse set of interest points. An ideal detector is able to find the corresponding int erest points even if the data undergo a transformation typical for a given domai n. Since the task is of high practical interest in computer vision, many hand-cr afted solutions were proposed. In this paper, we ask a fundamental question: can we learn such detectors from scratch? Since it is often unclear what points are "interesting", human labelling cannot be used to find a truly unbiased solution . Therefore, the task requires an unsupervised formulation. We are the first to propose such a formulation: training a neural network to rank points in a transf ormation-invariant manner. Interest points are then extracted from the top/botto m quantiles of this ranking. We validate our approach on two tasks: standard RGB image interest point detection and challenging cross-modal interest point detec tion between RGB and depth images. We quantitatively show that our unsupervised method performs better or on-par with baselines.

**********

Multi-Context Attention for Human Pose Estimation

Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, Xiaogang Wang; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1831-1840

In this paper, we propose to incorporate convolutional neural networks with a mu lti-context attention mechanism into an end-to-end framework for human pose esti mation. We adopt stacked hourglass networks to generate attention maps from feat ures at multiple resolutions with various semantics. The Conditional Random Fiel d (CRF) is utilized to model the correlations among neighboring regions in the a ttention map. We further combine the holistic attention model, which focuses on the global consistency of the full human body, and the body part attention model , which focuses on detailed descriptions for different body parts. Hence our mod el has the ability to focus on different granularity from local salient regions to global semantic consistent spaces. Additionally, we design novel Hourglass Re sidual Units (HRUs) to increase the receptive field of the network. These units are extensions of residual units with a side branch incorporating filters with l arger receptive field, hence features with various scales are learned and combin ed within the HRUs. The effectiveness of the proposed multi-context attention me chanism and the hourglass residual units is evaluated on two widely used human p ose estimation benchmarks. Our approach outperforms all existing methods on both benchmarks over all the body parts. Code has been made publicly available.

**********

Action Unit Detection With Region Adaptation, Multi-Labeling Learning and Optima l Temporal Fusing

Wei Li, Farnaz Abtahi, Zhigang Zhu; Proceedings of the IEEE Conference on Comput er Vision and Pattern Recognition (CVPR), 2017, pp. 1841-1850

Action Unit (AU) detection becomes essential for facial analysis. Many proposed approaches face challenging problems in dealing with the alignments of different

face regions, in the effective fusion of temporal information, and in training a model for multiple AU labels. To better address these problems, we propose a deep learning framework for AU detection with region of interest (ROI) adaptation, integrated multi-label learning, and optimal LSTM-based temporal fusing. First, an ROI cropping net is designed to make sure specific interested regions of faces are learned independently; each sub-region has a local convolutional neural network (CNN) whose convolutional filters will only be trained for the corresponding region. Second, multi-label learning is employed to integrate the outputs of those individual ROI cropping nets, which learns the inter-relationships of various AUs and acquires global features across sub-regions for AU detection. Finally, the optimal selection of multiple LSTM layers are carried out to best fuse temporal features, in order to make the AU prediction the most accurate. The proposed approach is evaluated on two popular AU detection datasets, BP4D and DISFA, outperforming the state of the art significantly, with an average improvement of around 13% in BP4D and 25% in DISFA, respectively.

**********

Unsupervised Learning of Depth and Ego-Motion From Video

Tinghui Zhou, Matthew Brown, Noah Snavely, David G. Lowe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1851-1858

We present an unsupervised learning framework for the task of dense 3D geometry and camera motion estimation from unstructured video sequences. In common with recent work, we use an end-to-end learning approach with view synthesis as the supervisory signal. In contrast to these works, our method is completely unsupervised, requiring only a sequence of images as input. We achieve this with a network that estimates the 6-DoF camera pose parameters of the input set, along with dense depth for a reference view using single-view inference. Our loss is constructed by projecting the nearby posed views into the reference view via the depth map. Results using the KITTI dataset demonstrate the effectiveness of our approach, which performs on par with another deep learning approach that assumes ground-truth pose information at training time.

**********

Joint Geometrical and Statistical Alignment for Visual Domain Adaptation

Jing Zhang, Wanqing Li, Philip Ogunbona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1859-1867

This paper presents a novel unsupervised domain adaptation method for cross-domain visual recognition. We propose a unified framework that reduces the shift between domains both statistically and geometrically, referred to as Joint Geometrical and Statistical Alignment (JGSA). Specifically, we learn two coupled projections that project the source domain and target domain data into low-dimensional subspaces where the geometrical shift and distribution shift are reduced simultaneously. The objective function can be solved efficiently in a closed form. Extensive experiments have verified that the proposed method significantly outperforms several state-of-the-art domain adaptation methods on a synthetic dataset and three different real world cross-domain visual recognition tasks.

**********

Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals

Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1868-1877

Physiological signals such as heart rate can provide valuable information about an individual's state and activity. However, existing work on computer vision has not yet explored leveraging these signals to enhance egocentric video understanding. In this work, we propose a model for reasoning on multimodal data to jointly predict activities and energy expenditures. We use heart rate signals as privileged self-supervision to derive energy expenditure in a training stage. A multitask objective is used to jointly optimize the two tasks. Additionally, we introduce a dataset that contains 31 hours of egocentric video augmented with heart rate and acceleration signals. This study can lead to new applications such as

a visual calorie counter.
**********

Attend in Groups: A Weakly-Supervised Deep Learning Framework for Learning From Web Data
Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1878-1887
Large-scale datasets have driven the rapid development of deep neural networks for visual recognition. However, annotating a massive dataset is expensive and time-consuming. Web images and their labels are, in comparison, much easier to obtain, but direct training on such automatially harvested images can lead to unsatisfactory performance, because the noisy labels of Web images adversely affect the learned recognition models. To address this drawback we propose an end-to-end weakly-supervised deep learning framework which is robust to the label noise in Web images. The proposed framework relies on two unified strategies -- random grouping and attention -- to effectively reduce the negative impact of noisy web image annotations. Specifically, random grouping stacks multiple images into a single training instance and thus increases the labeling accuracy at the instance level. Attention, on the other hand, suppresses the noisy signals from both incorrectly labeled images and less discriminative image regions. By conducting intensive experiments on two challenging datasets, including a newly collected fine-grained dataset with Web images of different car models, the superior performance of the proposed methods over competitive baselines is clearly demonstrated.
**********

An Exact Penalty Method for Locally Convergent Maximum Consensus
Huu Le, Tat-Jun Chin, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1888-1896
Maximum consensus estimation plays a critically important role in computer vision. Currently, the most prevalent approach draws from the class of non-deterministic hypothesize-and-verify algorithms, which are cheap but do not guarantee solution quality. On the other extreme, there are global algorithms which are exhaustive search in nature and can be costly for practical-sized inputs. This paper aims to fill the gap between the two extremes by proposing a locally convergent maximum consensus algorithm. Our method is based on a formulating the problem with linear complementarity constraints, then defining a penalized version which is provably equivalent to the original problem. Based on the penalty problem, we develop a Frank-Wolfe algorithm that can deterministically solve the maximum consensus problem. Compared to the randomized techniques, our method is deterministic and locally convergent; relative to the global algorithms, our method is much more practical on realistic input sizes. Further, our approach is naturally applicable to problems with geometric residuals.
**********

StyleBank: An Explicit Representation for Neural Image Style Transfer
Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1897-1906
We propose StyleBank, which is composed of multiple convolution filter banks and each filter bank explicitly represents one style, for neural image style transfer. To transfer an image to a specific style, the corresponding filter bank is operated on top of the intermediate feature embedding produced by a single auto-encoder. The StyleBank and the auto-encoder are jointly learnt, where the learning is conducted in such a way that the auto-encoder does not encode any style information thanks to the flexibility introduced by the explicit filter bank representation. It also enables us to conduct incremental learning to add a new image style by learning a new filter bank while holding the auto-encoder fixed. The explicit style representation along with the flexible network design enables us to fuse styles at not only the image level, but also the region level. Our method is the first style transfer network that links back to traditional texton mapping methods, and hence provides new understanding on neural style transfer. Our method is easy to train, runs in real-time, and produces results that qualitativel

y better or at least comparable to existing methods.
**********
Multi-View 3D Object Detection Network for Autonomous Driving
Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, Tian Xia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1907-1915
This paper aims at high-accuracy 3D object detection in autonomous driving scenario. We propose Multi-View 3D networks (MV3D), a sensory-fusion framework that takes both LIDAR point cloud and RGB images as input and predicts oriented 3D bounding boxes. We encode the sparse 3D point cloud with a compact multi-view representation. The network is composed of two subnetworks: one for 3D object proposal generation and another for multi-view feature fusion. The proposal network generates 3D candidate boxes efficiently from the bird's eye view representation of 3D point cloud. We design a deep fusion scheme to combine region-wise features from multiple views and enable interactions between intermediate layers of different paths. Experiments on the challenging KITTI benchmark show that our approach outperforms the state-of-the-art by around 25% and 30% AP on the tasks of 3D localization and 3D detection. In addition, for 2D detection, our approach obtains 14.9% higher AP than the state-of-the-art on the hard data among the LIDAR-based methods.
**********
Weakly Supervised Dense Video Captioning
Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, Xiangyang Xue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1916-1924
This paper focuses on a novel and challenging vision task, dense video captioning, which aims to automatically describe a video clip with multiple informative and diverse caption sentences. The proposed method is trained without explicit annotation of fine-grained sentence to video region-sequence correspondence, but is only based on weak video-level sentence annotations. It differs from existing video captioning systems in three technical aspects. First, we propose lexical fully convolutional neural networks (Lexical-FCN) with weakly supervised multi-instance multi-label learning to weakly link video regions with lexical labels. Second, we introduce a novel submodular maximization scheme to generate multiple informative and diverse region-sequences based on the Lexical-FCN outputs. A winner-takes-all scheme is adopted to weakly associate sentences to region-sequences in the training phase. Third, a sequence-to-sequence learning based language model is trained with the weakly supervised information obtained through the association process. We show that the proposed method can not only produce informative and diverse dense captions, but also outperform state-of-the-art single video captioning methods by a large margin.
**********
RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation
Guosheng Lin, Anton Milan, Chunhua Shen, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1925-1934
Recently, very deep convolutional neural networks (CNNs) have shown outstanding performance in object recognition and have also been the first choice for dense classification problems such as semantic segmentation. However, repeated subsampling operations like pooling or convolution striding in deep CNNs lead to a significant decrease in the initial image resolution. Here, we present RefineNet, a generic multi-path refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. In this way, the deeper layers that capture high-level semantic features can be directly refined using fine-grained features from earlier convolutions. The individual components of RefineNet employ residual connections following the identity mapping mindset, which allows for effective end-to-end training. Further, we introduce chained residual pooling, which captures rich background context in an efficient manner. We carry out comprehensive experiments and set new state-of-the-art results on seven public datasets. In particular, we achieve an intersection-over-union score of 83.4 on the chall

enging PASCAL VOC 2012 dataset, which is the best reported result to date.
**********

General Models for Rational Cameras and the Case of Two-Slit Projections
Matthew Trager, Bernd Sturmfels, John Canny, Martial Hebert, Jean Ponce; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1935-1943
The rational camera model recently introduced in [18] provides a general methodology for studying abstract nonlinear imaging systems and their multi-view geometry. This paper builds on this framework to study "physical realizations" of rational cameras. More precisely, we give an explicit account of the mapping between between physical visual rays and image points (missing in the original description), which allows us to give simple analytical expressions for direct and inverse projections. We also consider "primitive" camera models, that are orbits under the action of various projective transformations, and lead to a general notion of intrinsic parameters. The methodology is general, but it is illustrated concretely by an in-depth study of two-slit cameras, that we model using pairs of linear projections. This simple analytical form allows us to describe models for the corresponding primitive cameras, to introduce intrinsic parameters with a clear geometric meaning, and to define an epipolar tensor characterizing two-view correspondences. In turn, this leads to new algorithms for structure from motion and self-calibration.
**********

Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach
Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, Lizhen Qu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1944-1952
We present a theoretically grounded approach to train deep neural networks, including recurrent networks, subject to class-dependent label noise. We propose two procedures for loss correction that are agnostic to both application domain and network architecture. They simply amount to at most a matrix inversion and multiplication, provided that we know the probability of each class being corrupted into another. We further show how one can estimate these probabilities, adapting a recent technique for noise estimation to the multi-class setting, and thus providing an end-to-end framework. Extensive experiments on MNIST, IMDB, CIFAR-10, CIFAR-100 and a large scale dataset of clothing images employing a diversity of architectures --- stacking dense, convolutional, pooling, dropout, batch normalization, word embedding, LSTM and residual layers --- demonstrate the noise robustness of our proposals. Incidentally, we also prove that, when ReLU is the only non-linearity, the loss curvature is immune to class-dependent label noise.
**********

Semantic Segmentation via Structured Patch Prediction, Context CRF and Guidance CRF
Falong Shen, Rui Gan, Shuicheng Yan, Gang Zeng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1953-1961
This paper describes a fast and accurate semantic image segmentation approach that encodes not only segmentation-specified features but also high-order context compatibilities and boundary guidance constraints. We introduce a structured patch prediction technique to make a trade-off between classification discriminability and boundary sensibility for features. Both label and feature contexts are embedded to ensure recognition accuracy and compatibility, while the complexity of the high order cliques is reduced by a distance-aware sampling and pooling strategy. The proposed joint model also employs a guidance CRF to further enhance the segmentation performance. The message passing step is augmented with the guided filtering which enables an efficient and joint training of the whole system in an end-to-end fashion. Our proposed joint model outperforms the state-of-art on Pascal VOC 2012 and Cityscapes, with mIoU(%) of 82.5 and 79.2 respectively. It also reaches a leading performance on ADE20K, which is the dataset of the scene parsing track in ILSVRC 2016.
**********

Deep Matching Prior Network: Toward Tighter Multi-Oriented Text Detection

Yuliang Liu, Lianwen Jin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1962-1969
Detecting incidental scene text is a challenging task because of multi-orientation, perspective distortion, and variation of text size, color and scale. Retrospective research has only focused on using rectangular bounding box or horizontal sliding window to localize text, which may result in redundant background noise, unnecessary overlap or even information loss. To address these issues, we propose a new Convolutional Neural Networks (CNNs) based method, named Deep Matching Prior Network (DMPNet), to detect text with tighter quadrangle. First, we use quadrilateral sliding windows in several specific intermediate convolutional layers to roughly recall the text with higher overlapping area and then a shared Monte-Carlo method is proposed for fast and accurate computing of the polygonal areas. After that, we designed a sequential protocol for relative regression which can exactly predict text with compact quadrangle. Moreover, a auxiliary smooth Ln loss is also proposed for further regressing the position of text, which has better overall performance than L2 loss and smooth L1 loss in terms of robustness and stability. The effectiveness of our approach is evaluated on a public word-level, multi-oriented scene text database, ICDAR 2015 Robust Reading Competition Challenge 4 "Incidental scene text localization". The performance of our method is evaluated by using F-measure and found to be 70.64%, outperforming the existing state-of-the-art method with F-measure 63.76%.
**********

Person Search With Natural Language Description
Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1970-1979
Searching persons in large-scale image databases with the query of natural language description has important applications in video surveillance. Existing methods mainly focused on searching persons with image-based or attribute-based queries, which have major limitations for a practical usage. In this paper, we study the problem of person search with natural language description. Given the textual description of a person, the algorithm of the person search is required to rank all the samples in the person database then retrieve the most relevant sample corresponding to the queried description. Since there is no person dataset or benchmark with textual description available, we collect a large-scale person description dataset with detailed natural language annotations and person samples from various sources, termed as CUHK Person Description Dataset (CUHK-PEDES). A wide range of possible models and baselines have been evaluated and compared on the person search benchmark. An Recurrent Neural Network with Gated Neural Attention mechanism (GNA-RNN) is proposed to establish the state-of-the art performance on person search.
**********

Analyzing Computer Vision Data - The Good, the Bad and the Ugly
Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, Gustavo Fernandez Dominguez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1980-1990
In recent years, a great number of datasets were published to train and evaluate computer vision (CV) algorithms. These valuable contributions helped to push CV solutions to a level where they can be used for safety-relevant applications, such as autonomous driving. However, major questions concerning quality and usefulness of test data for CV evaluation are still unanswered. Researchers and engineers try to cover all test cases by using as much test data as possible. In this paper, we propose a different solution for this challenge. We introduce a method for dataset analysis which builds upon an improved version of the CV-HAZOP checklist, a list of potential hazards within the CV domain. Picking stereo vision as an example, we provide an extensive survey of 28 datasets covering the last two decades. We create a tailored checklist and apply it to the datasets Middlebury, KITTI, Sintel, Freiburg, and HCI to present a thorough characterization and quantitative comparison. We confirm the usability of our checklist for identification of challenging stereo situations by applying nine state-of-the-art stereo

matching algorithms on the analyzed datasets, showing that hazard frames correla
te with difficult frames. We show that challenging datasets still allow a meanin
gful algorithm evaluation even for small subsets. Finally, we provide a list of
missing test cases that are still not covered by current datasets as inspiration
 for researchers who want to participate in future dataset creation.
**********

3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation F
rom Single Depth Images
Liuhao Ge, Hui Liang, Junsong Yuan, Daniel Thalmann; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1991-2000
We propose a simple, yet effective approach for real-time hand pose estimation f
rom single depth images using three-dimensional Convolutional Neural Networks (3
D CNNs). Image based features extracted by 2D CNNs are not directly suitable for
 3D hand pose estimation due to the lack of 3D spatial information. Our proposed
 3D CNN taking a 3D volumetric representation of the hand depth image as input c
an capture the 3D spatial structure of the input and accurately regress full 3D
hand pose in a single pass. In order to make the 3D CNN robust to variations in
hand sizes and global orientations, we perform 3D data augmentation on the train
ing data. Experiments show that our proposed 3D CNN based approach outperforms s
tate-of-the-art methods on two challenging hand pose datasets, and is very effic
ient as our implementation runs at over 215 fps on a standard computer with a si
ngle GPU.
**********

iCaRL: Incremental Classifier and Representation Learning
Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, Christoph H. Lamper
t; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
 (CVPR), 2017, pp. 2001-2010
A major open problem on the road to artificial intelligence is the development o
f incrementally learning systems that learn about more and more concepts over ti
me from a stream of data. In this work, we introduce a new training strategy, iC
aRL, that allows learning in such a class-incremental way: only the training dat
a for a small number of classes has to be present at the same time and new class
es can be added progressively. iCaRL learns strong classifiers and a data repres
entation simultaneously. This distinguishes it from earlier works that were fund
amentally limited to fixed data representations and therefore incompatible with
deep learning architectures. We show by experiments on CIFAR-100 and ImageNet IL
SVRC 2012 data that iCaRL can learn many classes incrementally over a long perio
d of time where other strategies quickly fail.
**********

PoseTrack: Joint Multi-Person Pose Estimation and Tracking
Umar Iqbal, Anton Milan, Juergen Gall; Proceedings of the IEEE Conference on Com
puter Vision and Pattern Recognition (CVPR), 2017, pp. 2011-2020
In this work, we introduce the challenging problem of joint multi-person pose es
timation and tracking of an unknown number of persons in unconstrained videos. E
xisting methods for multi-person pose estimation in images cannot be applied dir
ectly to this problem, since it also requires to solve the problem of person ass
ociation over time in addition to the pose estimation for each person. We theref
ore propose a novel method that jointly models multi-person pose estimation and
tracking in a single formulation. To this end, we represent body joint detection
s in a video by a spatio-temporal graph and solve an integer linear program to p
artition the graph into sub-graphs that correspond to plausible body pose trajec
tories for each person. The proposed approach implicitly handles occlusion and t
runcation of persons. Since the problem has not been addressed quantitatively in
 the literature, we introduce a challenging "Multi-Person PoseTrack" dataset, an
d also propose a completely unconstrained evaluation protocol that does not make
 any assumptions about the scale, size, location or the number of persons. Final
ly, we evaluate the proposed approach and several baseline methods on our new da
taset.
**********

Learning a Deep Embedding Model for Zero-Shot Learning

Li Zhang, Tao Xiang, Shaogang Gong; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2017, pp. 2021-2030
Zero-shot learning (ZSL) models rely on learning a joint embedding space where b
oth textual/semantic description of object classes and visual representation of
object images can be projected to for nearest neighbour search. Despite the succ
ess of deep neural networks that learn an end-to-end model between text and imag
es in other vision problems such as image captioning, very few deep ZSL model ex
ists and they show little advantage over ZSL models that utilise deep feature re
presentations but do not learn an end-to-end embedding. In this paper we argue t
hat the key to make deep ZSL models succeed is to choose the right embedding spa
ce. Instead of embedding into a semantic space or an intermediate space, we prop
ose to use the visual space as the embedding space. This is because that in this
 space, the subsequent nearest neighbour search would suffer much less from the
hubness problem and thus become more effective. This  model design also provides
 a natural mechanism for multiple semantic modalities (e.g., attributes and sent
ence descriptions) to be fused and optimised jointly in an end-to-end manner. Ex
tensive experiments on four benchmarks show that our model significantly outperf
orms the existing models.
**********

MCMLSD: A Dynamic Programming Approach to Line Segment Detection
Emilio J. Almazan, Ron Tal, Yiming Qian, James H. Elder; Proceedings of the IEEE
 Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2031-20
39
Prior approaches to line segment detection typically involve perceptual grouping
 in the image domain or global accumulation in the Hough domain.  Here we propos
e a probabilistic algorithm that merges the advantages of both approaches.  In a
 first stage lines are detected using a global probabilistic Hough approach.  In
 the second stage each detected line is analyzed in the image domain to localize
 the line segments that generated the peak in the Hough map.  By limiting search
 to a line, the distribution of segments over the sequence of points on the line
 can be modeled as a Markov chain, and a probabilistically optimal labelling can
 be computed exactly using a standard dynamic programming algorithm, in linear t
ime.  The Markov assumption also leads to an intuitive ranking method that uses
the local marginal posterior probabilities to estimate the expected number of co
rrectly labelled points on a segment.  To assess the resulting Markov Chain Marg
inal Line Segment Detector (MCMLSD) we develop and apply a novel quantitative ev
aluation methodology that controls for under- and over-segmentation.  Evaluation
 on the YorkUrbanDB dataset shows that the proposed MCMLSD method outperforms th
e state-of-the-art by a substantial margin.
**********

Deep MANTA: A Coarse-To-Fine Many-Task Network for Joint 2D and 3D Vehicle Analy
sis From Monocular Image
Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Celine Teuliere, Thierry Cha
teau; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2017, pp. 2040-2049
In this paper, we present a novel approach, called Deep MANTA (Deep Many-Tasks),
 for many-task vehicle analysis from a given image. A robust convolutional netwo
rk is introduced for simultaneous vehicle detection, part localization, visibili
ty characterization and 3D dimension estimation. Its architecture is based on a
new coarse-to-fine object proposal that boosts the vehicle detection. Moreover,
the Deep MANTA network is able to localize vehicle parts even if these parts are
 not visible. In the inference, the network's outputs are used by a real time ro
bust pose estimation algorithm for fine orientation estimation and 3D vehicle lo
calization. We show in experiments that our method outperforms monocular state-o
f-the-art approaches on vehicle detection, orientation and 3D location tasks on
the very challenging KITTI benchmark.
**********

Low-Rank Embedded Ensemble Semantic Dictionary for Zero-Shot Learning
Zhengming Ding, Ming Shao, Yun Fu; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2017, pp. 2050-2058

Zero-shot learning for visual recognition has received much interest in the most recent years. However, the semantic gap across visual features and their underlying semantics is still the biggest obstacle in zero-shot learning. To fight off this hurdle, we propose an effective Low-rank Embedded Semantic Dictionary learning (LESD) through ensemble strategy. Specifically, we formulate a novel framework to jointly seek a low-rank embedding and semantic dictionary to link visual features with their semantic representations, which manages to capture shared features across different observed classes. Moreover, ensemble strategy is adopted to learn multiple semantic dictionaries to constitute the latent basis for the unseen classes. Consequently, our model could extract a variety of visual characteristics within objects, which can be well generalized to unknown categories. Extensive experiments on several zero-shot benchmarks verify that the proposed model can outperform the state-of-the-art approaches.
**********

Nonnegative Matrix Underapproximation for Robust Multiple Model Fitting
Mariano Tepper, Guillermo Sapiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2059-2067
In this work, we introduce a highly efficient algorithm to address the nonnegative matrix underapproximation (NMU) problem, i.e., nonnegative matrix factorization (NMF) with an additional underapproximation constraint. NMU results are interesting as, compared to traditional NMF, they present additional sparsity and part-based behavior, explaining unique data features. To show these features in practice, we first present an application to the analysis of climate data. We then present an NMU-based algorithm to robustly fit multiple parametric models to a dataset. The proposed approach delivers state-of-the-art results for the estimation of multiple fundamental matrices and homographies, outperforming other alternatives in the literature and exemplifying the use of efficient NMU computations.
**********

BIND: Binary Integrated Net Descriptors for Texture-Less Object Recognition
Jacob Chan, Jimmy Addison Lee, Qian Kemao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2068-2076
This paper presents BIND (Binary Integrated Net Descriptor), a texture-less object detector that encodes multi-layered binary-represented nets for high precision edge-based description. Our proposed concept aligns layers of object-sized patches (nets) onto highly fragmented occlusion resistant line-segment midpoints (linelets) to encode regional information into efficient binary strings. These lightweight nets encourage discriminative object description through their high-spatial resolution, enabling highly precise encoding of the object's edges and internal texture-less information. BIND achieved various invariant properties such as rotation, scale and edge-polarity through its unique binary logical-operated encoding and matching techniques, while performing remarkably well in occlusion and clutter. Apart from yielding efficient computational performance, BIND also attained remarkable recognition rates surpassing recent state-of-the-art texture-less object detectors such as BORDER, BOLD and LINE2D.
**********

Efficient Diffusion on Region Manifolds: Recovering Small Objects With Compact CNN Representations
Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, Ondrej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2077-2086
Query expansion is a popular method to improve the quality of image retrieval with both conventional and CNN representations. It has been so far limited to global image similarity. This work focuses on diffusion, a mechanism that captures the image manifold in the feature space. An efficient off-line stage allows optional reduction in the number of stored regions. In the on-line stage, the proposed handling of unseen queries in the indexing stage removes additional computation to adjust the precomputed data. We perform diffusion through a sparse linear system solver, yielding practical query times well below one second. Experimentally, we observe a significant boost in performance of image retrieval with compact CNN descriptors on standard benchmarks, especially when the query object cove

rs only a small part of the image. Small objects have been a common failure case of CNN-based retrieval.

**********

S2F: Slow-To-Fast Interpolator Flow

Yanchao Yang, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2087-2096

We introduce a method to compute optical flow at multiple scales of motion, without resorting to multi- resolution or combinatorial methods. It addresses the key problem of small objects moving fast, and resolves the artificial binding between how large an object is and how fast it can move before being diffused away by clas- sical scale-space. Even with no learning, it achieves top performance on the most challenging optical flow benchmark. Moreover, the results are interpretable, and indeed we list the assumptions underlying our method explicitly. The key to our approach is the matching pro- gression from slow to fast, as well as the choice of in- terpolation method, or equivalently the prior, to fill in regions where the data allows it. We use several off- the-shelf components, with relatively low sensitivity to parameter tuning. Computational cost is comparable to the state-of-the-art.

**********

ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2097-2106

The chest X-ray is one of the most commonly accessible radiological examinations for screening and diagnosis of many lung diseases. A tremendous number of X-ray imaging studies accompanied by radiological reports are accumulated and stored in many modern hospitals' Picture Archiving and Communication Systems (PACS). On the other side, it is still an open question how this type of hospital-size knowledge database containing invaluable imaging informatics (i.e., loosely labeled) can be used to facilitate the data-hungry deep learning paradigms in building truly large-scale high precision computer-aided diagnosis (CAD) systems. In this paper, we present a new chest X-ray database, namely "ChestX-ray8", which comprises 108,948 frontal-view X-ray images of 32,717 unique patients with the text-mined eight disease image labels (where each image can have multi-labels), from the associated radiological reports using natural language processing. Importantly, we demonstrate that these commonly occurring thoracic diseases can be detected and even spatially-located via a unified weakly-supervised multi-label image classification and disease localization framework, which is validated using our proposed dataset. Although the initial quantitative results are promising as reported, deep convolutional neural network based "reading chest X-rays" (i.e., recognizing and locating the common disease patterns trained with only image-level labels) remains a strenuous task for fully-automated high precision CAD systems.

**********

Learning From Simulated and Unsupervised Images Through Adversarial Training

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, Russell Webb; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2107-2116

With recent progress in graphics, it has become more tractable to train models on synthetic images, potentially avoiding the need for expensive annotations. However, learning from synthetic images may not achieve the desired performance due to a gap between synthetic and real image distributions. To reduce this gap, we propose Simulated+Unsupervised (S+U) learning, where the task is to learn a model to improve the realism of a simulator's output using unlabeled real data, while preserving the annotation information from the simulator. We develop a method for S+U learning that uses an adversarial network similar to Generative Adversarial Networks (GANs), but with synthetic images as inputs instead of random vectors. We make several key modifications to the standard GAN algorithm to preserve annotations, avoid artifacts, and stabilize training: (i) a 'self-regularizatio

n' term, (ii) a local adversarial loss, and (iii) updating the discriminator usi ng a history of refined images. We show that this enables generation of highly r ealistic images, which we demonstrate both qualitatively and with a user study. We quantitatively evaluate the generated images by training models for gaze esti mation and hand pose estimation. We show a significant improvement over using sy nthetic images, and achieve state-of-the-art results on the MPIIGaze dataset wit hout any labeled real data.

**********

Feature Pyramid Networks for Object Detection

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Reco gnition (CVPR), 2017, pp. 2117-2125

Feature pyramids are a basic component in recognition systems for detecting obje cts at different scales. But pyramid representations have been avoided in recent object detectors that are based on deep convolutional networks, partially becau se they are slow to compute and memory intensive. In this paper, we exploit the inherent multi-scale, pyramidal hierarchy of deep convolutional networks to cons truct feature pyramids with marginal extra cost. A top-down architecture with la teral connections is developed for building high-level semantic feature maps at all scales. This architecture, called a Feature Pyramid Network (FPN), shows sig nificant improvement as a generic feature extractor in several applications. Usi ng a basic Faster R-CNN system, our method achieves state-of-the-art single-mode l results on the COCO detection benchmark without bells and whistles, surpassing all existing single-model entries including those from the COCO 2016 challenge winners. In addition, our method can run at 5 FPS on a GPU and thus is a practic al and accurate solution to multi-scale object detection. Code will be made publ icly available.

**********

Loss Max-Pooling for Semantic Image Segmentation

Samuel Rota Bulo, Gerhard Neuhold, Peter Kontschieder; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2126-2135

In this work we introduce a novel loss max-pooling concept for handling imbalanc ed training data distributions, applicable as alternative loss layer in the cont ext of deep neural networks for semantic image segmentation tasks. Most real-wor ld semantic segmentation datasets exhibit long tail distributions with few objec t categories comprising the majority of data and consequently biasing the classi fiers towards them. Our method adaptively re-weights the contributions of each p ixel based on their observed losses, targeting under-performing classification r esults as often encountered for under-represented object classes. Moreover, our approach goes beyond conventional cost-sensitive learning attempts through adapt ive considerations that allow us to indirectly address both, inter- and intra-cl ass imbalances. We provide a theoretical justification of our approach, compleme ntary to experimental analyses on standard semantic segmentation datasets. In ou r experiments on the challenging Cityscapes and Pascal VOC 2012 segmentation ben chmarks we find consistently improved results, demonstrating the efficacy of our approach.

**********

Learned Contextual Feature Reweighting for Image Geo-Localization

Hyo Jin Kim, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2136-2145

We address the problem of large scale image geo-localization where the location of an image is estimated by identifying geo-tagged reference images depicting th e same place. We propose a novel model for learning image representations that i ntegrates context-aware feature reweighting in order to effectively focus on reg ions that positively contribute to geo-localization. In particular, we introduce a Contextual Reweighting Network (CRN) that predicts the importance of each reg ion in the feature map based on the image context. Our model is learned end-to-e nd for the image geo-localization task, and requires no annotation other than im age geo-tags for training. In experimental results, the proposed approach signif icantly outperforms the previous state-of-the-art on the standard geo-localizati

on benchmark datasets. We also demonstrate that our CRN discovers task-relevant contexts without any additional supervision.
**********

On the Effectiveness of Visible Watermarks
Tali Dekel, Michael Rubinstein, Ce Liu, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2146-2154
Visible watermarking is a widely-used technique for marking and protecting copyrights of many millions of images on the web, yet it suffers from an inherent security flaw---watermarks are typically added in a consistent manner to many images. We show that this consistency allows to automatically estimate the watermark and recover the original images with high accuracy. Specifically, we present a generalized multi-image matting algorithm that takes a watermarked image collection as input and automatically estimates the "foreground" (watermark), its alpha matte, and the "background" (original) images. Since such an attack relies on the consistency of watermarks across image collection, we explore and evaluate how it is affected by various types of inconsistencies in the watermark embedding that could potentially be used to make watermarking more secured. We demonstrate the algorithm on stock imagery available on the web, and provide extensive quantitative analysis on synthetic watermarked data. A key takeaway message of this paper is that visible watermarks should be designed to not only be robust against removal from a single image, but to be more resistant to mass-scale removal from image collections as well.
**********

Deep View Morphing
Dinghuang Ji, Junghyun Kwon, Max McFarland, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2155-2163
Recently, convolutional neural networks (CNN) have been successfully applied to view synthesis problems. However, such CNN-based methods can suffer from lack of texture details, shape distortions, or high computational complexity. In this paper, we propose a novel CNN architecture for view synthesis called "Deep View Morphing" that does not suffer from these issues. To synthesize a middle view of two input images, a rectification network first rectifies the two input images. An encoder-decoder network then generates dense correspondences between the rectified images and blending masks to predict the visibility of pixels of the rectified images in the middle view. A view morphing network finally synthesizes the middle view using the dense correspondences and blending masks. We experimentally show the proposed method significantly outperforms the state-of-the-art CNN-based view synthesis method.
**********

Designing Illuminant Spectral Power Distributions for Surface Classification
Henryk Blasinski, Joyce Farrell, Brian Wandell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2164-2173
There are many scientific, medical and industrial imaging applications where users have full control of the scene illumination and color reproduction is not the primary objective For example, it is possible to co-design sensors and spectral illumination in order to classify and detect changes in biological tissues, organic and inorganic materials, and object surface properties. In this paper, we propose two different approaches to illuminant spectrum selection for surface classification. In the supervised framework we formulate a biconvex optimization problem where we alternate between optimizing support vector classifier weights and optimal illuminants. We also describe a sparse Principal Component Analysis (PCA) dimensionality reduction approach that can be used with unlabeled data. We efficiently solve the non-convex PCA problem using a convex relaxation and Alternating Direction Method of Multipliers (ADMM). We compare the classification accuracy of a monochrome imaging sensor with optimized illuminants to the classification accuracy of conventional RGB cameras with natural broadband illumination.
**********

End-To-End Learning of Driving Models From Large-Scale Video Datasets

Huazhe Xu, Yang Gao, Fisher Yu, Trevor Darrell; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2174-2182
Robust perception-action models should be learned from training data with divers e visual appearances and realistic behaviors, yet current approaches to deep vis uomotor policy learning have been generally limited to in-situ models learned fr om a single vehicle or simulation environment. We advocate learning a generic ve hicle motion model from large scale crowd-sourced video data, and develop an end -to-end trainable architecture for learning to predict a distribution over futur e vehicle egomotion from instantaneous monocular camera observations and previou s vehicle state. Our model incorporates a novel FCN-LSTM architecture, which can be learned from large-scale crowd-sourced vehicle action data, and leverages av ailable scene segmentation side tasks to improve performance under a privileged learning paradigm. We provide a novel large-scale dataset of crowd-sourced drivi ng behavior suitable for training our model, and report results predicting the d river action on held out sequences across diverse conditions.
**********

Unsupervised Visual-Linguistic Reference Resolution in Instructional Videos
De-An Huang, Joseph J. Lim, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 218 3-2192
We propose an unsupervised method for reference resolution in instructional vide os, where the goal is to temporally link an entity (e.g., "dressing") to the act ion (e.g., "mix yogurt") that produced it. The key challenge is the inevitable v isual-linguistic ambiguities arising from the changes in both visual appearance and referring expression of an entity in the video. This challenge is amplified by the fact that we aim to resolve references with no supervision. We address th ese challenges by learning a joint visual-linguistic model, where linguistic cue s can help resolve visual ambiguities and vice versa. We verify our approach by learning our model unsupervisedly using more than two thousand unstructured cook ing videos from YouTube, and show that our visual-linguistic model can substanti ally improve upon state-of-the-art linguistic only model on reference resolution in instructional videos.
**********

Dense Captioning With Joint Inference and Visual Context
Linjie Yang, Kevin Tang, Jianchao Yang, Li-Jia Li; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2193-2202
Dense captioning is a newly emerging computer vision topic for understanding ima ges with dense language descriptions. The goal is to densely detect visual conce pts (e.g., objects, object parts, and interactions between them) from images, la beling each with a short descriptive phrase. We identify two key challenges of d ense captioning that need to be properly addressed when tackling the problem. Fi rst, dense visual concept annotations in each image are associated with highly o verlapping target regions, making accurate localization of each visual concept c hallenging. Second, the large amount of visual concepts makes it hard to recogni ze each of them by appearance alone. We propose a new model pipeline based on tw o novel ideas, joint inference and context fusion, to alleviate these two challe nges. We design our model architecture in a methodical manner and thoroughly eva luate the variations in architecture. Our final model, compact and efficient, a chieves state-of-the-art accuracy on Visual Genome for dense captioning with a r elative gain of 73% compared to the previous best algorithm. Qualitative experim ents also reveal the semantic capabilities of our model in dense captioning.
**********

Unsupervised Learning of Long-Term Motion Dynamics for Videos
Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, Li Fei-Fei; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2203-2212
We present an unsupervised representation learning approach that compactly encod es the motion dependencies in videos. Given a pair of images from a video clip, our framework learns to predict the long-term 3D motions. To reduce the complexi ty of the learning framework, we propose to describe the motion as a sequence of

atomic 3D flows computed with RGB-D modality. We use a Recurrent Neural Networ
k based Encoder-Decoder framework to predict these sequences of flows. We argue
that in order for the decoder to reconstruct these sequences, the encoder must l
earn a robust video representation that captures long-term motion dependencies a
nd spatial-temporal relations. We demonstrate the effectiveness of our learned t
emporal representations on activity classification across multiple modalities a
nd datasets such as NTU RGB+D and MSR Daily Activity 3D. Our framework is gener
ic to any input modality, i.e., RGB, depth, and RGB-D videos.
**********

CLKN: Cascaded Lucas-Kanade Networks for Image Alignment
Che-Han Chang, Chun-Nan Chou, Edward Y. Chang; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2213-2221
This paper proposes a data-driven approach for image alignment. Our main contrib
ution is a novel network architecture that combines the strengths of convolution
al neural networks (CNNs) and the Lucas-Kanade algorithm. The main component of
this architecture is a Lucas-Kanade layer that performs the inverse compositiona
l algorithm on convolutional feature maps. To train our network, we develop a ca
scaded feature learning method that incorporates the coarse-to-fine strategy int
o the training process. This method learns a pyramid representation of convoluti
onal features in a cascaded manner and yields a cascaded network that performs c
oarse-to-fine alignment on the feature pyramids. We apply our model to the task
of homography estimation, and perform training and evaluation on a large labeled
dataset generated from the MS-COCO dataset. Experimental results show that the
proposed approach significantly outperforms the other methods.
**********

Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localizati
on
Kuo-Hao Zeng, Shih-Han Chou, Fu-Hsiang Chan, Juan Carlos Niebles, Min Sun; Proce
edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
2017, pp. 2222-2230
For survival, a living agent (e.g., human in Fig. 1(a)) must have the ability to
assess risk (1) by temporally anticipating accidents before they occur (Fig. 1(
b)), and (2) by spatially localizing risky regions (Fig. 1(c)) in the environmen
t to move away from threats. In this paper, we take an agent-centric approach to
study the accident anticipation and risky region localization tasks. We propose
a novel soft-attention Recurrent Neural Network (RNN) which explicitly models b
oth spatial and appearance-wise non-linear interaction between the agent trigger
ing the event and another agent or static-region involved. In order to test our
proposed method, we introduce the Epic Fail (EF) dataset consisting of 3000 vira
l videos capturing various accidents. In the experiments, we evaluate the risk a
ssessment accuracy both in the temporal domain (accident anticipation) and spati
al domain (risky region localization) on our EF dataset and the Street Accident
(SA) dataset. Our method consistently outperforms other baselines on both datase
ts.
**********

ShapeOdds: Variational Bayesian Learning of Generative Shape Models
Shireen Elhabian, Ross Whitaker; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2017, pp. 2231-2242
Shape models provide a compact parameterization of a class of shapes, and have b
een shown to be important to a variety of vision problems, including object dete
ction, tracking, and image segmentation. Learning generative shape models from g
rid-structured representations, aka silhouettes, is usually hindered by (1) data
likelihoods with intractable marginals and posteriors, (2) high-dimensional sha
pe spaces with limited training samples (and the associated risk of overfitting)
, and (3) estimation of hyperparameters relating to model complexity that often
entails computationally expensive grid searches. In this paper, we propose a Bay
esian treatment that relies on direct probabilistic formulation for learning gen
erative shape models in the silhouettes space. We propose a variational approach
for learning a latent variable model in which we make use of, and extend, recen
t works on variational bounds of logistic-Gaussian integrals to circumvent intra

ctable marginals and posteriors. Spatial coherency and sparsity priors are also incorporated to lend stability to the optimization problem by regularizing the s olution space while avoiding overfitting in this high-dimensional, low-sample-si ze scenario. We deploy a type-II maximum likelihood estimate of the model hyperp arameters to avoid grid searches. We demonstrate that the proposed model generat es realistic samples, generalizes to unseen examples, and is able to handle miss ing regions and/or background clutter, while comparing favorably with recent, ne ural-network-based approaches.
**********

Expecting the Unexpected: Training Detectors for Unusual Pedestrians With Advers arial Imposters
Shiyu Huang, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2243-2252
As autonomous vehicles become an every-day reality, high-accuracy pedestrian det ection is of paramount practical importance.  Pedestrian detection is a highly r esearched topic with mature methods,  but most datasets (for both training and e valuation) focus on common scenes of people engaged in typical walking poses on sidewalks. But performance is most crucial for dangerous scenarios that are rare ly observed, such as children playing in the street and people using bicycles/sk ateboards in unexpected ways. Such "in-the-tail" data is notoriously hard to obs erve, making both training and testing difficult. To analyze this problem, we ha ve collected a novel annotated dataset of dangerous scenarios called the Precari ous Pedestrian dataset. Even given a  dedicated collection effort, it is relativ ely small by contemporary standards (~ 1000 images). To explore large-scale data -driven learning, we explore the use of synthetic data generated by a game engin e. A significant challenge is selected the right "priors" or parameters for synt hesis: we would like realistic data with realistic poses and object configuratio ns. Inspired by Generative Adversarial Networks, we generate a massive amount of synthetic data and train a discriminative classifier to select a realistic subs et (that fools the classifier), which we deem Synthetic Imposters. We demonstrat e that this pipeline allows one to generate realistic (or adverserial) training data by making use of rendering/animation engines. Interestingly, we also demons trate that such data can be used to rank algorithms, suggesting that Synthetic I mposters can also be used for "in-the-tail" validation at test-time, a notorious ly difficult challenge for real-world deployment.
**********

ER3: A Unified Framework for Event Retrieval, Recognition and Recounting
Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, Nann ing Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Rec ognition (CVPR), 2017, pp. 2253-2262
We develop a unified framework for complex event retrieval, recognition and reco unting. The framework is based on a compact video representation that exploits t he temporal correlations in image features. Our feature alignment procedure iden tifies and removes the feature redundancies across frames and outputs an interme diate tensor representation we call video imprint. The video imprint is then fed into a reasoning network, whose attention mechanism parallels that of memory ne tworks used in language modeling. The reasoning network simultaneously recognize s the event category and locates the key pieces of evidence for event recounting . In event retrieval tasks, we show that the compact video representation aggreg ated from the video imprint achieves significantly better retrieval accuracy com pared with existing methods. We also set new state of the art results in event r ecognition tasks with an additional benefit: The latent structure in our reasoni ng network highlights the areas of the video imprint and can be directly used fo r event recounting. As video imprint maps back to locations in the video frames, the network allows not only the identification of key frames but also specific areas inside each frame which are most influential to the decision process.
**********

Outlier-Robust Tensor PCA
Pan Zhou, Jiashi Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2263-2271

Low-rank tensor analysis is important for various real applications in computer vision. However, existing methods focus on recovering a low-rank tensor contaminated by Gaussian or gross sparse noise and hence cannot effectively handle outliers that are common in practical tensor data. To solve this issue, we propose an outlier-robust tensor principle component analysis (OR-TPCA) method for simultaneous low-rank tensor recovery and outlier detection. For intrinsically low-rank tensor observations with arbitrary outlier corruption, OR-TPCA is the first method that has provable performance guarantee for exactly recovering the tensor subspace and detecting outliers under mild conditions. Since tensor data are naturally high-dimensional and multi-way, we further develop a fast randomized algorithm that requires small sampling size yet can substantially accelerate OR-TPCA without performance drop. Experimental results on four tasks: outlier detection, clustering, semi-supervised and supervised learning, clearly demonstrate the advantages of our method.

**********

Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, Wangmeng Zuo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2272-2281

In domain adaptation, maximum mean discrepancy (MMD) has been widely adopted as a discrepancy metric between the distributions of source and target domains. However, existing MMD-based domain adaptation methods generally ignore the changes of class prior distributions, i.e., class weight bias across domains. This remains an open problem but ubiquitous for domain adaptation, which can be caused by changes in sample selection criteria and application scenarios. We show that MMD cannot account for class weight bias and results in degraded domain adaptation performance. To address this issue, a weighted MMD model is proposed in this paper. Specifically, we introduce class-specific auxiliary weights into the original MMD for exploiting the class prior probability on source and target domains, whose challenge lies in the fact that the class label in target domain is unavailable. To account for it, our proposed weighted MMD model is defined by introducing an auxiliary weight for each class in the source domain, and a classification EM algorithm is suggested by alternating between assigning the pseudo-labels, estimating auxiliary weights and updating model parameters. Extensive experiments demonstrate the superiority of our weighted MMD over conventional MMD for domain adaptation.

**********

SyncSpecCNN: Synchronized Spectral CNN for 3D Shape Segmentation

Li Yi, Hao Su, Xingwen Guo, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2282-2290

In this paper, we study the problem of semantic annotation on 3D models that are represented as shape graphs. A functional view is taken to represent localized information on graphs, so that annotations such as part segment or keypoint are nothing but 0-1 indicator vertex functions. Compared with images that are 2D grids, shape graphs are irregular and non-isomorphic data structures. To enable the prediction of vertex functions on them by convolutional neural networks, we resort to spectral CNN method that enables weight sharing by parametrizing kernels in the spectral domain spanned by graph Laplacian eigenbases. Under this setting, our network, named SyncSpecCNN, strives to overcome two key challenges: how to share coefficients and conduct multi-scale analysis in different parts of the graph for a single shape, and how to share information across related but different shapes that may be represented by very different graphs. Towards these goals, we introduce a spectral parametrization of dilated convolutional kernels and a spectral transformer network. Experimentally we tested SyncSpecCNN on various tasks, including 3D shape part segmentation and keypoint prediction. State-of-the-art performance has been achieved on all benchmark datasets.

**********

Unrolling the Shutter: CNN to Correct Motion Distortions

Vijay Rengarajan, Yogesh Balaji, A. N. Rajagopalan; Proceedings of the IEEE Conf

erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2291-2299
Row-wise exposure delay present in CMOS cameras is responsible for skew and curv ature distortions known as the rolling shutter (RS) effect while imaging under c amera motion. Existing RS correction methods resort to using multiple images or tailor scene-specific correction schemes. We propose a convolutional neural netw ork (CNN) architecture that automatically learns essential scene features from a single RS image to estimate the row-wise camera motion and undo RS distortions back to the time of first-row exposure. We employ long rectangular kernels to sp ecifically learn the effects produced by the row-wise exposure. Experiments reve al that our proposed architecture performs better than the conventional CNN empl oying square kernels. Our single-image correction method fares well even operati ng in a frame-by-frame manner against video-based methods and performs better th an scene-specific correction schemes even under challenging situations.
**********

Deep Level Sets for Salient Object Detection
Ping Hu, Bing Shuai, Jun Liu, Gang Wang; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2017, pp. 2300-2309
Deep learning has been applied to saliency detection in recent years. The superi or performance has proved that deep networks can model the semantic properties o f salient objects. Yet it is difficult for a deep network to discriminate pixels belonging to similar receptive fields around the object boundaries, thus deep n etworks may output maps with blurred saliency and inaccurate boundaries. To tack le such an issue, in this work, we propose a deep Level Set network to produce c ompact and uniform saliency maps. Our method drives the network to learn a Level Set function for salient objects so it can output more accurate boundaries and compact saliency. Besides, to propagate saliency information among pixels and re cover full resolution saliency map, we extend a superpixel-based guided filter t o be a layer in the network. The proposed network has a simple structure and is trained end-to-end. During testing, the network can produce saliency maps by eff iciently feedforwarding testing images at a speed over 12FPS on GPUs. Evaluation s on benchmark datasets show that the proposed method achieves state-of-the-art performance.
**********

Instance-Aware Image and Sentence Matching With Selective Multimodal LSTM
Yan Huang, Wei Wang, Liang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2310-2318
Effective image and sentence matching depends on how to well measure their globa l visual-semantic similarity. Based on the observation that such a global simila rity arises from a complex aggregation of multiple local similarities between pa irwise instances of image (objects) and sentence (words), we propose a selective multimodal Long Short-Term Memory network (sm-LSTM) for instance-aware image an d sentence matching. The sm-LSTM includes a multimodal context-modulated attenti on scheme at each timestep that can selectively attend to a pair of instances of image and sentence, by predicting pairwise instance-aware saliency maps for ima ge and sentence. For selected pairwise instances, their representations are obta ined based on the predicted saliency maps, and then compared to measure their lo cal similarity. By similarly measuring multiple local similarities within a few timesteps, the sm-LSTM sequentially aggregates them with hidden states to obtain a final matching score as the desired global similarity. Extensive experiments show that our model can well match image and sentence with complex content, and achieve the state-of-the-art results on two public benchmark datasets.
**********

From Motion Blur to Motion Flow: A Deep Learning Solution for Removing Heterogen eous Motion Blur
Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton van den Hengel, Qinfeng Shi; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2017, pp. 2319-2328
Removing pixel-wise heterogeneous motion blur is challenging due to the ill-pose d nature of the problem. The predominant solution is to estimate the blur kernel by adding a prior, but extensive literature on the subject indicates the diffic

ulty in identifying a prior which is suitably informative, and general. Rather t
han imposing a prior based on theory, we propose instead to learn one from the d
ata. Learning a prior over the latent image would require modeling all possible
image content. The critical observation underpinning our approach, however, is t
hat learning the motion flow instead allows the model to focus on the cause of t
he blur, irrespective of the image content. This is a much easier learning task,
but it also avoids the iterative process through which latent image priors are
typically applied. Our approach directly estimates the motion flow from the blur
red image through a fully-convolutional deep neural network (FCN) and recovers t
he unblurred image from the estimated motion flow. Our FCN is the first universa
l end-to-end mapping from the blurred image to the dense motion flow. To train t
he FCN, we simulate motion flows to generate synthetic blurred-image-motion-flow
pairs thus avoiding the need for human labeling. Extensive experiments on chall
enging realistic blurred images demonstrate that the proposed method outperforms
the state-of-the-art.
**********

Deep Temporal Linear Encoding Networks
Ali Diba, Vivek Sharma, Luc Van Gool; Proceedings of the IEEE Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2017, pp. 2329-2338
The CNN-encoding of features from entire videos for the representation of human
actions has rarely been addressed. Instead, CNN work has focused on approaches t
o fuse spatial and temporal networks, but these were typically limited to proces
sing shorter sequences. We present a new video representation, called temporal l
inear encoding (TLE) and embedded inside of CNNs as a new layer, which captures
the appearance and motion throughout entire videos. It encodes this aggregated i
nformation into a robust video feature representation, via end-to-end learning.
Advantages of TLEs are: (a) they encode the entire video into a compact feature
representation, learning the semantics and a discriminative feature space; (b) t
hey are applicable to all kinds of networks like 2D and 3D CNNs for video classi
fication; and (c) they model feature interactions in a more expressive way and w
ithout loss of information. We conduct experiments on two challenging human acti
on datasets: HMDB51 and UCF101. The experiments show that TLE outperforms curren
t state-of-the-art methods on both datasets.
**********

End-To-End Training of Hybrid CNN-CRF Models for Stereo
Patrick Knobelreiter, Christian Reinbacher, Alexander Shekhovtsov, Thomas Pock;
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C
VPR), 2017, pp. 2339-2348
We propose a novel and principled hybrid CNN+CRF model for stereo estimation. Ou
r model allows to exploit the advantages of both, convolutional neural networks
(CNNs) and conditional random fields (CRFs) in an unified approach. The CNNs com
pute expressive features for matching and distinctive color edges, which in turn
are used to compute the unary and binary costs of the CRF. For inference, we ap
ply a recently proposed highly parallel dual block descent algorithm which only
needs a small fixed number of iterations to compute a high-quality approximate m
inimizer. As the main contribution of the paper, we propose a theoretically soun
d method based on the structured output support vector machine (SSVM) to train t
he hybrid CNN+CRF model on large-scale data end-to-end. Our trained models perfo
rm very well despite the fact that we are using shallow CNNs and do not apply an
y kind of post-processing to the final output of the CRF. We evaluate our combin
ed models on challenging stereo benchmarks such as Middlebury 2014 and Kitti 201
5 and also investigate the performance of each individual component.
**********

Deep Feature Flow for Video Recognition
Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, Yichen Wei; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2349-2
358
Deep convolutional neutral networks have achieved great success on image recogni
tion tasks. Yet, it is non-trivial to transfer the state-of-the-art image recogn
ition networks to videos as per-frame evaluation is too slow and unaffordable. W

e present deep feature flow, a fast and accurate framework for video recognition. It runs the expensive convolutional sub-network only on sparse key frames and propagates their deep feature maps to other frames via a flow field. It achieves significant speedup as flow computation is relatively fast. The end-to-end training of the whole architecture significantly boosts the recognition accuracy. Deep feature flow is flexible and general. It is validated on two recent large scale video datasets. It makes a large step towards practical video recognition. Code would be released.

**********

Fully Convolutional Instance-Aware Semantic Segmentation

Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, Yichen Wei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2359-2367

We present the first fully convolutional end-to-end solution for instance-aware semantic segmentation task. It inherits all the merits of FCNs for semantic segmentation and instance mask proposal. It performs instance mask prediction and classification jointly. The underlying convolutional representation is fully shared between the two sub-tasks, as well as between all regions of interest. The network architecture is highly integrated and efficient. It achieves state-of-the-art performance in both accuracy and efficiency. It wins the COCO 2016 segmentation competition by a large margin. Code would be released.

**********

Truncated Max-Of-Convex Models

Pankaj Pansari, M. Pawan Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2368-2376

Truncated convex models (TCM) are a special case of pair- wise random fields that have been widely used in computer vision. However, by restricting the order of the potentials to be at most two, they fail to capture useful image statistics. We propose a natural generalization of TCM to high-order random fields, which we call truncated max-of-convex models (TMCM). The energy function of TMCM consists of two types of potentials: (i) unary potential, which has no restriction on its form; and (ii) high-order potential, which is the sum of the truncation of the m largest convex distances over disjoint pairs of random variables in an arbitrary size clique. The use of a convex distance function encourages smoothness, while truncation allows for discontinuities in the labeling. By using m > 1, TMCM provides robustness towards errors in the definition of the cliques. In order to minimize the energy function of a TMCM over all possible labelings, we design an efficient st-mincut based range expansion algorithm. We prove the accuracy of our algorithm by establishing strong multiplicative bounds for several special cases of interest. Using synthetic and standard real datasets, we demonstrate the benefit of our high-order TMCM over pairwise TCM, as well as the benefit of our range expansion algorithm over other st-mincut based approaches.

**********

Asymmetric Feature Maps With Application to Sketch Based Retrieval

Giorgos Tolias, Ondrej Chum; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2377-2385

We propose a novel concept of asymmetric feature maps (AFM), which allows to evaluate multiple kernels between a query and database entries without increasing the memory requirements. To demonstrate the advantages of the AFM method, we derive a short vector image representation that, due to asymmetric feature maps, supports efficient scale and translation invariant sketch-based image retrieval. Unlike most of the short-code based retrieval systems, the proposed method provides the query localization in the retrieved image. The efficiency of the search is boosted by approximating a 2D translation search via trigonometric polynomial of scores by 1D projections. The projections are a special case of AFM. An order of magnitude speed-up is achieved compared to traditional trigonometric polynomials. The results are boosted by an image-based average query expansion, exceeding significantly the state of the art on standard benchmarks.

**********

Instance-Level Salient Object Segmentation

Guanbin Li, Yuan Xie, Liang Lin, Yizhou Yu; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2386-2395
Image saliency detection has recently witnessed rapid progress due to deep convo
lutional neural networks. However, none of the existing methods is able to ident
ify object instances in the detected salient regions. In this paper, we present
a salient instance segmentation method that produces a saliency mask with distin
ct object instance labels for an input image. Our method consists of three steps
, estimating saliency map, detecting salient object contours and identifying sal
ient object instances. For the first two steps, we propose a multiscale saliency
 refinement network, which generates high-quality salient region masks and salie
nt object contours. Once integrated with multiscale combinatorial grouping and a
 MAP-based subset optimization framework, our method can generate very promising
 salient object instance segmentation results. To promote further research and e
valuation of salient instance segmentation, we also construct a new database of
1000 images and their pixelwise salient instance annotations. Experimental resul
ts demonstrate that our proposed method is capable of achieving state-of-the-art
 performance on all public benchmarks for salient region detection as well as on
 our new dataset for salient instance segmentation.
**********

Kernel Square-Loss Exemplar Machines for Image Retrieval
Rafael S. Rezende, Joaquin Zepeda, Jean Ponce, Francis Bach, Patrick Perez; Proc
eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
, 2017, pp. 2396-2404
Zepeda and Perez have recently demonstrated the promise of the exemplar SVM (ESV
M) as a feature encoder for image retrieval. This paper extends this approach in
 several directions: We first show that replacing the hinge loss by the square l
oss in the ESVM cost function significantly reduces encoding time with negligibl
e effect on accuracy. We call this model square-loss exemplar machine, or SLEM.
 We then introduce a kernelized SLEM which can be implemented efficiently throug
h low-rank matrix decomposition, and displays improved performance.  Both SLEM v
ariants exploit the fact that the negative examples are fixed, so most of the SL
EM computational complexity is relegated to an offline process independent of th
e positive examples.  Our experiments establish the performance and computationa
l advantages of our approach using a large array of base features and standard i
mage retrieval datasets.
**********

Direct Photometric Alignment by Mesh Deformation
Kaimo Lin, Nianjuan Jiang, Shuaicheng Liu, Loong-Fah Cheong, Minh Do, Jiangbo Lu
; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
(CVPR), 2017, pp. 2405-2413
The choice of motion models is vital in applications like image/video stitching
and video stabilization. Conventional methods explored different approaches rang
ing from simple global parametric models to complex per-pixel optical flow. Mesh
-based warping methods achieve a good balance between computational complexity a
nd model flexibility. However, they typically require high quality feature corre
spondences and suffer from mismatches and low-textured image content. In this pa
per, we propose a mesh-based photometric alignment method that minimizes pixel i
ntensity difference instead of Euclidean distance of known feature correspondenc
es. The proposed method combines the superior performance of dense photometric a
lignment with the efficiency of mesh-based image warping. It achieves better glo
bal alignment quality than the feature-based counterpart in textured images, and
 more importantly, it is also robust to low-textured image content. Abundant exp
eriments show that our method can handle a variety of images and videos, and out
performs representative state-of-the-art methods in both image stitching and vid
eo stabilization tasks.
**********

Semantic Multi-View Stereo: Jointly Estimating Objects and Voxels
Ali Osman Ulusoy, Michael J. Black, Andreas Geiger; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2414-2423
Dense 3D reconstruction from RGB images is a highly ill-posed problem due to occ

lusions, textureless or reflective surfaces, as well as other challenges. We propose object-level shape priors to address these ambiguities. Towards this goal, we formulate a probabilistic model that integrates multi-view image evidence with 3D shape information from multiple objects. Inference in this model yields a dense 3D reconstruction of the scene as well as the existence and precise 3D pose of the objects in it. Our approach is able to recover fine details not captured in the input shapes while defaulting to the input models in occluded regions where image evidence is weak. Due to its probabilistic nature, the approach is able to cope with the approximate geometry of the 3D models as well as input shapes that are not present in the scene. We evaluate the approach quantitatively on several challenging indoor and outdoor datasets.
**********

HOPE: Hierarchical Object Prototype Encoding for Efficient Object Instance Search in Videos
Tan Yu, Yuwei Wu, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2424-2433
This paper tackles the problem of efficient and effective object instance search in videos. To effectively capture the relevance between a query and video frames and precisely localize the particular object, we leverage the object proposals to improve the quality of object instance search in videos. However, hundreds of object proposals obtained from each frame could result in unaffordable memory and computational cost. To this end, we present a simple yet effective hierarchical object prototype encoding (HOPE) model to accelerate the object instance search without sacrificing accuracy, which exploits both the spatial and temporal self-similarity property existing in object proposals generated from video frames. We design two types of sphere k-means methods, i.e., spatially-constrained sphere k-means and temporally-constrained sphere k-means to learn frame-level object prototypes and dataset-level object prototypes, respectively.  In this way, the object instance search problem is cast to the sparse matrix-vector multiplication problem. Thanks to the sparsity of the codes, both the memory and computational cost are significantly reduced. Experimental results on two video datasets demonstrate that our approach significantly improves the performance of video object instance search over other state-of-the-art fast search schemes.
**********

Learning Adaptive Receptive Fields for Deep Image Parsing Network
Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, Si Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2434-2442
In this paper, we introduce a novel approach to regulate receptive field in deep image parsing network automatically. Unlike previous works which have stressed much importance on obtaining better receptive fields using manually selected dilated convolutional kernels, our approach uses two affine transformation layers in the network's backbone and operates on feature maps. Feature maps will be inflated/shrinked by the new layer and therefore receptive fields in following layers are changed accordingly. By end-to-end training, the whole framework is data-driven without laborious manual intervention. The proposed method is generic across dataset and different tasks. We conduct extensive experiments on both general parsing task and face parsing task as concrete examples to demonstrate the method's superior regulation ability over manual designs.
**********

Contour-Constrained Superpixels for Image and Video Processing
Se-Ho Lee, Won-Dong Jang, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2443-2451
A novel contour-constrained superpixel (CCS) algorithm is proposed in this work. We initialize superpixels and regions in a regular grid and then refine the superpixel label of each region hierarchically from block to pixel levels. To make superpixel boundaries compatible with object contours, we propose the notion of contour pattern matching and formulate an objective function including the contour constraint. Furthermore, we extend the CCS algorithm to generate temporal superpixels for video processing. We initialize superpixel labels in each frame by transferring those in the previous frame and refine the labels to make superpixe

ls temporally consistent as well as compatible with object contours. Experimenta l results demonstrate that the proposed algorithm provides better performance th an the state-of-the-art superpixel methods.
**********

Learning to Predict Stereo Reliability Enforcing Local Consistency of Confidence Maps
Matteo Poggi, Stefano Mattoccia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2452-2461
Confidence measures estimate unreliable disparity assignments performed by a ste reo matching algorithm and, as recently proved, can be used for several purposes . This paper aims at increasing, by means of a deep network, the effectiveness o f state-of-the-art confidence measures exploiting the local consistency assumpti on. We exhaustively evaluated our proposal on 23 confidence measures, including 5 top-performing ones based on random-forests and CNNs, training our networks wi th two popular stereo algorithms and a small subset (25 out of 194 frames) of th e KITTI 2012 dataset. Experimental results show that our approach dramatically i ncreases the effectiveness of all the 23 confidence measures on the remaining fr ames. Moreover, without re-training, we report a further cross-evaluation on KIT TI 2015 and Middlebury 2014 confirming that our proposal provides remarkable imp rovements for each confidence measure even when dealing with significantly diffe rent input data. To the best of our knowledge, this is the first method to move beyond conventional pixel-wise confidence estimation.
**********

FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks
Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, Tho mas Brox; Proceedings of the IEEE Conference on Computer Vision and Pattern Reco gnition (CVPR), 2017, pp. 2462-2470
The FlowNet demonstrated that optical flow estimation can be cast as a learning problem. However, the state of the art with regard to the quality of the flow ha s still been defined by traditional methods. Particularly on small displacements and real-world data, FlowNet cannot compete with variational methods. In this p aper, we advance the concept of end-to-end learning of optical flow and make it work really well. The large improvements in quality and speed are caused by thre e major contributions: first, we focus on the training data and show that the sc hedule of presenting data during training is very important. Second, we develop a stacked architecture that includes warping of the second image with intermedia te optical flow. Third, we elaborate on small displacements by introducing a sub network specializing on small motions. FlowNet 2.0 is only marginally slower tha n the original FlowNet but decreases the estimation error by more than 50%. It p erforms on par with state-of-the-art methods, while running at interactive frame rates. Moreover, we present faster variants that allow optical flow computation at up to 140fps with accuracy matching the original FlowNet.
**********

Growing a Brain: Fine-Tuning by Increasing Model Capacity
Yu-Xiong Wang, Deva Ramanan, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2471-2480
CNNs have made an undeniable impact on computer vision through the ability to le arn high-capacity models with large annotated training sets. One of their remark able properties is the ability to transfer knowledge from a large source dataset to a (typically smaller) target dataset. This is usually accomplished through f ine-tuning a fixed-size network on new target data. Indeed, virtually every cont emporary visual recognition system makes use of fine-tuning to transfer knowledg e from ImageNet. In this work, we analyze what components and parameters change during fine-tuning, and discover that increasing model capacity allows for more natural model adaptation through fine-tuning. By making an analogy to developmen tal learning, we demonstrate that growing a CNN with additional units, either by widening existing layers or deepening the overall network, significantly outper forms classic fine-tuning approaches. But in order to properly grow a network, w e show that newly-added units must be appropriately normalized to allow for a pa ce of learning that is consistent with existing units. We empirically validate o

ur approach on several benchmark datasets, producing state-of-the-art results.
**********

Dynamic Attention-Controlled Cascaded Shape Regression Exploiting Training Data Augmentation and Fuzzy-Set Sample Weighting
Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, Xiao-Jun Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2481-2490
We present a new Cascaded Shape Regression (CSR) architecture, namely Dynamic Attention-Controlled CSR (DAC-CSR), for robust facial landmark detection on unconstrained faces. Our DAC-CSR divides facial landmark detection into three cascaded sub-tasks: face bounding box refinement, general CSR and attention-controlled CSR. The first two stages refine initial face bounding boxes and output intermediate facial landmarks. Then, an online dynamic model selection method is used to choose appropriate domain-specific CSRs for further landmark refinement. The key innovation of our DAC-CSR is the fault-tolerant mechanism, using fuzzy set sample weighting, for attention-controlled domain-specific model training. Moreover, we advocate data augmentation with a simple but effective 2D profile face generator, and context-aware feature extraction for better facial feature representation. Experimental results obtained on challenging datasets demonstrate the merits of our DAC-CSR over the state-of-the-art methods.
**********

Additive Component Analysis
Calvin Murdock, Fernando De la Torre; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2491-2499
Principal component analysis (PCA) is one of the most versatile tools for unsupervised learning with applications ranging from dimensionality reduction to exploratory data analysis and visualization. While much effort has been devoted to encouraging meaningful representations through regularization (e.g. non-negativity or sparsity), underlying linearity assumptions can limit their effectiveness. To address this issue, we propose Additive Component Analysis (ACA), a novel nonlinear extension of PCA. Inspired by multivariate nonparametric regression with additive models, ACA fits a smooth manifold to data by learning an explicit mapping from a low-dimensional latent space to the input space, which trivially enables applications like denoising. Furthermore, ACA can be used as a drop-in replacement in many algorithms that use linear component analysis methods as a subroutine via the local tangent space of the learned manifold. Unlike many other nonlinear dimensionality reduction techniques, ACA can be efficiently applied to large datasets since it does not require computing pairwise similarities or storing training data during testing. Multiple ACA layers can also be composed and learned jointly with essentially the same procedure for improved representational power, demonstrating the encouraging potential of nonparametric deep learning. We evaluate ACA on a variety of datasets, showing improved robustness, reconstruction performance, and interpretability.
**********

Lifting From the Deep: Convolutional 3D Pose Estimation From a Single Image
Denis Tome, Chris Russell, Lourdes Agapito; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2500-2509
We propose a unified formulation for the problem of 3D human pose estimation from a single raw RGB image that reasons jointly about 2D joint estimation and 3D pose reconstruction to improve both tasks. We take an integrated approach that fuses probabilistic knowledge of 3D human pose with a multi-stage CNN architecture and uses the knowledge of plausible 3D landmark locations to refine the search for better 2D locations. The entire process is trained end-to-end, is extremely efficient and obtains state-of-the-art results on Human3.6M outperforming previous approaches both on 2D and 3D errors.
**********

Attentional Push: A Deep Convolutional Network for Augmenting Image Salience With Shared Attention Modeling in Social Scenes
Siavash Gorji, James J. Clark; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2510-2519

We present a novel visual attention tracking technique based on Shared Attention modeling. By considering the viewer as a participant in the activity occurring in the scene, our model learns the loci of attention of the scene actors and use it to augment image salience. We go beyond image salience and instead of only computing the power of image regions to pull attention, we also consider the strength with which the scene actors push attention to the region in question, thus the term Attentional Push. We present a convolutional neural network (CNN) which augments standard saliency models with Attentional Push. Our model contains two pathways: an Attentional Push pathway which learns the gaze location of the scene actors and a saliency pathway. These are followed by a shallow augmented saliency CNN which combines them and generates the augmented saliency. For training, we use transfer learning to initialize and train the Attentional Push CNN by minimizing the classification error of following the actors' gaze location on a 2-D grid using a large-scale gaze-following dataset. The Attentional Push CNN is then fine-tuned along with the augmented saliency CNN to minimize the Euclidean distance between the augmented saliency and ground truth fixations using an eye-tracking dataset, annotated with the head and the gaze location of the scene actors. We evaluate our model on three challenging eye fixation datasets, SALICON, iSUN and CAT2000, and illustrate significant improvements in predicting viewers' fixations in social scenes.

**********

Fine-Grained Recognition as HSnet Search for Informative Image Parts
Michael Lam, Behrooz Mahasseni, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2520-2529

This work addresses fine-grained image classification. Our work is based on the hypothesis that when dealing with subtle differences among object classes it is critical to identify and only account for a few informative image parts, as the remaining image context may not only be uninformative but may also hurt recognition. This motivates us to formulate our problem as a sequential search for informative parts over a deep feature map produced by a deep Convolutional Neural Network (CNN). A state of this search is a set of proposal bounding boxes in the image, whose "informativeness" is evaluated by the heuristic function (H), and used for generating new candidate states by the successor function (S). The two functions are unified via a Long Short-Term Memory network (LSTM) into a new deep recurrent architecture, called HSnet. Thus, HSnet (i) generates proposals of informative image parts and (ii) fuses all proposals toward final fine-grained recognition. We specify both supervised and weakly supervised training of HSnet depending on the availability of object part annotations. Evaluation on the benchmark Caltech-UCSD Birds 200-2011 and Cars-196 datasets demonstrate our competitive performance relative to the state of the art.

**********

Scalable Person Re-Identification on Supervised Smoothed Manifold
Song Bai, Xiang Bai, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2530-2539

Most existing person re-identification algorithms either extract robust visual features or learn discriminative metrics for person images. However, the underlying manifold which those images reside on is rarely investigated. That arises a problem that the learned metric is not smooth with respect to the local geometry structure of the data manifold.  In this paper, we study person re-identification with manifold-based affinity learning, which did not receive enough attention from this area. An unconventional manifold-preserving algorithm is proposed, which can 1) make best use of supervision from training data, whose label information is given as pairwise constraints; 2) scale up to large repositories with low on-line time complexity; and 3) be plunged into most existing algorithms, serving as a generic postprocessing procedure to further boost the identification accuracies. Extensive experimental results on five popular person re-identification benchmarks consistently demonstrate the effectiveness of our method. Especially, on the largest CUHK03 and Market-1501, our method outperforms the state-of-the-art alternatives by a large margin with high efficiency, which is more appropriate for practical applications.

**********
Riemannian Nonlinear Mixed Effects Models: Analyzing Longitudinal Deformations in Neuroimaging
Hyunwoo J. Kim, Nagesh Adluru, Heemanshu Suri, Baba C. Vemuri, Sterling C. Johnson, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2540-2549
Statistical machine learning models that operate on manifold-valued data are being extensively studied in vision, motivated by applications in activity recognition, feature tracking and medical imaging. While non-parametric methods have been relatively well studied in the literature, efficient formulations for parametric models (which may offer benefits in small sample size regimes) have only emerged recently. So far, manifold-valued regression models (such as geodesic regression) are restricted to the analysis of cross-sectional data, i.e., the so-called "fixed effects" in statistics. But in most "longitudinal analysis" (e.g., when a participant provides multiple measurements, over time) the application of fixed effects models is problematic. In an effort to answer this need, this paper generalizes non-linear mixed effects model to the regime where the response variable is manifold-valued, i.e., f:R^d -> M. We derive the underlying model and estimation schemes and demonstrate the immediate benefits such a model can provide --- both for group level and individual level analysis --- on longitudinal brain imaging data. The direct consequence of our results is that longitudinal analysis of manifold-valued measurements (especially, the symmetric positive definite manifold) can be conducted in a computationally tractable manner.
**********
Detecting Oriented Text in Natural Images by Linking Segments
Baoguang Shi, Xiang Bai, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2550-2558
Most state-of-the-art text detection methods are specific to horizontal Latin text and are not fast enough for real-time applications. We introduce Segment Linking (SegLink), an oriented text detection method. The main idea is to decompose text into two locally detectable elements, namely segments and links. A segment is an oriented box covering a part of a word or text line; A link connects two adjacent segments, indicating that they belong to the same word or text line. Both elements are detected densely at multiple scales by an end-to-end trained, fully-convolutional neural network. Final detections are produced by combining segments connected by links. Compared with previous methods, SegLink improves along the dimensions of accuracy, speed, and ease of training. It achieves an f-measure of 75.0% on the standard ICDAR 2015 Incidental (Challenge 4) benchmark, outperforming the previous best by a large margin. It runs at over 20 FPS on 512x512 images. Moreover, without modification, SegLink is able to detect long lines of non-Latin text, such as Chinese.
**********
Diverse Image Annotation
Baoyuan Wu, Fan Jia, Wei Liu, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2559-2567
In this work we study the task of image annotation, of which the goal is to describe an image using a few tags. Instead of predicting the full list of tags, here we target for providing a short list of tags under a limited number (e.g., 3), to cover as much information as possible of the image. The tags in such a short list should be representative and diverse. It means they are required to be not only corresponding to the contents of the image, but also be different to each other. To this end, we treat the image annotation as a subset selection problem based on the conditional determinantal point process (DPP) model, which formulates the representation and diversity jointly. We further explore the semantic hierarchy and synonyms among the candidate tags, and require that two tags in a semantic hierarchy or in a pair of synonyms should not be selected simultaneously. This requirement is then embedded into the sampling algorithm according to the learned conditional DPP model. Besides, we find that traditional metrics for image annotation (e.g., precision, recall and F1 score) only consider the representation, but ignore the diversity. Thus we propose new metrics to evaluate the

quality of the selected subset (i.e., the tag list), based on the semantic hierarchy and synonyms. Human study through Amazon Mechanical Turk verifies that the proposed metrics are more close to the human's judgment than traditional metrics. Experiments on two benchmark datasets show that the proposed method can produce more representative and diverse tags, compared with existing image annotation methods.
**********

Inverse Compositional Spatial Transformer Networks
Chen-Hsuan Lin, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2568-2576
In this paper, we establish a theoretical connection between the classical Lucas & Kanade (LK) algorithm and the emerging topic of Spatial Transformer Networks (STNs). STNs are of interest to the vision and learning communities due to their natural ability to combine alignment and classification within the same theoretical framework. Inspired by the Inverse Compositional (IC) variant of the LK algorithm, we present Inverse Compositional Spatial Transformer Networks (IC-STNs). We demonstrate that IC-STNs can achieve better performance than conventional STNs with less model capacity; in particular, we show superior performance in pure image alignment tasks as well as joint alignment/classification problems on real-world problems.
**********

Factorized Variational Autoencoders for Modeling Audience Reactions to Movies
Zhiwei Deng, Rajitha Navarathna, Peter Carr, Stephan Mandt, Yisong Yue, Iain Matthews, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2577-2586
Matrix and tensor factorization methods are often used for finding underlying low-dimensional patterns from noisy data. In this paper, we study non-linear tensor factoriza- tion methods based on deep variational autoencoders. Our approach is well-suited for settings where the relationship between the latent representation to be learned and the raw data representation is highly complex. We apply our ap- proach to a large dataset of facial expressions of movie- watching audiences (over 16 million faces). Our experi- ments show that compared to conventional linear factoriza- tion methods, our method achieves better reconstruction of the data, and further discovers interpretable latent factors.
**********

Adversarially Tuned Scene Generation
VSR Veeravasarapu, Constantin Rothkopf, Ramesh Visvanathan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2587-2595
Generalization performance of trained computer vision (CV) systems that use computer graphics (CG) generated data is not yet effective due to the concept of 'domain-shift' between virtual and real data. Although simulated data augmented with a few real-world samples has been shown to mitigate domain shift and improve transferability of trained models, guiding or bootstrapping the virtual data generation with the distributions learnt from target real world domain is desired, especially in the fields where annotating even few real images is laborious (such as semantic labeling, optical flow, and intrinsic images etc.). In order to address this problem in an unsupervised manner, our work combines recent advances in CG, which aims at generating stochastic scene layouts using large collections of 3D object models, and generative adversarial training, which aims at training generative models by measuring discrepancy between generated and real data in terms of their separability in the space of a deep discriminatively-trained classifier. Our method uses iterative estimation of the posterior density of prior distributions for a generative graphical model. This is done within a rejection sampling framework. Initially, we assume uniform distributions as priors over parameters of a scene described by a generative graphical model. As iterations proceed the uniform prior distributions are updated sequentially to distributions that are closer to the unknown distributions of target data. We demonstrate the utility of adversarially tuned scene generation on two real world benchmark datasets (CityScapes and CamVid) for traffic scene semantic labeling with a d

eep convolutional net (DeepLab). We obtained performance improvements by 2.28 an
d 3.14 points on the IoU metric between the DeepLab models trained on simulated
sets prepared from the scene generation models before and after tuning to CitySc
apes and CamVid respectively.
**********

Binge Watching: Scaling Affordance Learning From Sitcoms
Xiaolong Wang, Rohit Girdhar, Abhinav Gupta; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2596-2605
In recent years, there has been a renewed interest in jointly modeling perceptio
n and action. At the core of this investigation is the idea of modeling affordan
ces. However, when it comes to predicting affordances, even the state of the art
 approaches still do not use any ConvNets. Why is that? Unlike semantic or 3D ta
sks, there still does not exist any large-scale dataset for affordances. In this
 paper, we tackle the challenge of creating one of the biggest dataset for learn
ing affordances. We use seven sitcoms to extract a diverse set of scenes and how
 actors interact with different objects in the scenes. Our dataset consists of m
ore than 10K scenes and 28K ways humans can interact with these 10K images. We a
lso propose a two-step approach to predict affordances in a new scene. In the fi
rst step, given a location in the scene we classify which of the 30 pose classes
 is the likely affordance pose. Given the pose class and the scene, we then use
a Variational Autoencoder (VAE) to extract the scale and deformation of the pose
. The VAE allows us to sample the distribution of possible poses at test time. F
inally, we show the importance of large-scale data in learning a generalizable a
nd robust model of affordances.
**********

A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection
Xiaolong Wang, Abhinav Shrivastava, Abhinav Gupta; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2606-2615
How do we learn an object detector that is invariant to occlusions and deformati
ons? Our current solution is to use a data-driven strategy -- collect large-scal
e datasets which have object instances under different conditions. The hope is t
hat the final classifier can use these examples to learn invariances. But is it
really possible to see all the occlusions in a dataset? We argue that like categ
ories, occlusions and object deformations also follow a long-tail. Some occlusio
ns and deformations are so rare that they  hardly happen; yet we want to learn a
 model invariant to such occurrences. In this paper, we propose an alternative s
olution. We propose to learn an adversarial network that generates examples with
 occlusions and deformations. The goal of the adversary is to generate examples
that are difficult for the object detector to classify. In our framework both th
e original detector and adversary are learned in a  joint manner. Our experiment
al results indicate a 2.3% mAP boost on VOC07 and a 2.6% mAP boost on VOC2012 ob
ject detection challenge compared to the Fast-RCNN pipeline.
**********

Cognitive Mapping and Planning for Visual Navigation
Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, Jitendra Malik;
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C
VPR), 2017, pp. 2616-2625
We introduce a neural architecture for navigation in novel environments. Our pro
posed architecture learns to map from first-person views and plans a sequence of
 actions towards goals in the environment. The Cognitive Mapper and Planner (CMP
) is based on two key ideas: a) a unified joint architecture for mapping and pla
nning, such that the mapping is driven by the needs of the planner, and b) a spa
tial memory with the ability to plan given an incomplete set of observations abo
ut the world. CMP constructs a top-down belief map of the world and applies a di
fferentiable neural net planner to produce the next action at each time step. Th
e accumulated belief of the world enables the agent to track visited regions of
the environment. Our experiments demonstrate that CMP outperforms both reactive
strategies and standard memory-based architectures and performs well in novel en
vironments. Furthermore, we show that CMP can also achieve semantically specifie
d goals, such as "go to a chair".

**********
Multi-View Supervision for Single-View Reconstruction via Differentiable Ray Consistency
Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2626-2634
We study the notion of consistency between a 3D shape and a 2D observation and propose a differentiable formulation which allows computing gradients of the 3D shape given an observation from an arbitrary view. We do so by reformulating view consistency using a differentiable ray consistency (DRC) term. We show that this formulation can be incorporated in a learning framework to leverage different types of multi-view observations e.g. foreground masks, depth, color images, semantics etc. as supervision for learning single-view 3D prediction. We present empirical analysis of our technique in a controlled setting. We also show that this approach allows us to improve over existing techniques for single-view reconstruction of objects from the PASCAL VOC dataset.
**********
Learning Shape Abstractions by Assembling Volumetric Primitives
Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2635-2643
We present a learning framework for abstracting complex shapes by learning to assemble objects using 3D volumetric primitives. In addition to generating simple and geometrically interpretable explanations of 3D objects, our framework also allows us to automatically discover and exploit consistent structure in the data. We demonstrate that using our method allows predicting shape representations which can be leveraged for obtaining a consistent parsing across the instances of a shape collection and constructing an interpretable shape similarity measure. We also examine applications for image-based prediction as well as shape manipulation.
**********
AMC: Attention guided Multi-modal Correlation Learning for Image Search
Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2644-2652
Given a user's query, traditional image search systems rank images according to its relevance to a single modality (e.g., image content or surrounding text). Nowadays, an increasing number of images on the Internet are available with associated meta data in rich modalities (e.g., titles, keywords, tags, etc.), which can be exploited for better similarity measure with queries. In this paper, we leverage visual and textual modalities for image search by learning their correlation with input query. According to the intent of query, attention mechanism can be introduced to adaptively balance the importance of different modalities. We propose a novel Attention guided Multi-modal Correlation (AMC) learning method which consists of a jointly learned hierarchy of intra and inter-attention networks. Conditioned on query's intent, intra-attention networks (i.e., visual intra-attention network and language intra-attention network) attend on informative parts within each modality; a multi-modal inter-attention network promotes the importance of the most query-relevant modalities. In experiments, we evaluate AMC models on the search logs from two real world image search engines and show a significant boost on the ranking of user-clicked images in search results. Additionally, we extend AMC models to caption ranking task on COCO dataset and achieve competitive results compared with recent state-of-the-arts.
**********
Bidirectional Multirate Reconstruction for Temporal Modeling in Videos
Linchao Zhu, Zhongwen Xu, Yi Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2653-2662
Despite the recent success of neural networks in image feature learning, a major problem in the video domain is the lack of sufficient labeled data for learning to model temporal information. In this paper, we propose an unsupervised tempor

al modeling method that learns from untrimmed videos. The speed of motion varies constantly, e.g., a man may run quickly or slowly. We therefore train a Multirate Visual Recurrent Model (MVRM) by encoding frames of a clip with different intervals. This learning process makes the learned model more capable of dealing with motion speed variance. Given a clip sampled from a video, we use its past and future neighboring clips as the temporal context, and reconstruct the two temporal transitions, i.e., present->past transition and present->future transition, reflecting the temporal information in different views. The proposed method exploits the two transitions simultaneously by incorporating a bidirectional reconstruction which consists of a backward reconstruction and a forward reconstruction. We apply the proposed method to two challenging video tasks, i.e., complex event detection and video captioning, in which it achieves state-of-the-art performance. Notably, our method generates the best single feature for event detection with a relative improvement of 10.4% on the MEDTest-13 dataset and achieves the best performance in video captioning across all evaluation metrics on the YouTube2Text dataset.

**********

Learning Video Object Segmentation From Static Images

Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, Alexander Sorkine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2663-2672

Inspired by recent advances of deep learning in instance segmentation and object tracking, we introduce the concept of convnet-based guidance applied to video object segmentation. Our model proceeds on a per-frame basis, guided by the output of the previous frame towards the object of interest in the next frame. We demonstrate that highly accurate object segmentation in videos can be enabled by using a convolutional neural network (convnet) trained with static images only. The key component of our approach is a combination of offline and online learning strategies, where the former produces a refined mask from the previous' frame estimate and the latter allows to capture the appearance of the specific object instance. Our method can handle different types of input annotations such as bounding boxes and segments while leveraging an arbitrary amount of annotated frames. Therefore our system is suitable for diverse applications with different requirements in terms of accuracy and efficiency. In our extensive evaluation, we obtain competitive results on three different datasets, independently from the type of input annotation.

**********

The More You Know: Using Knowledge Graphs for Image Classification

Kenneth Marino, Ruslan Salakhutdinov, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2673-2681

One characteristic that sets humans apart from modern learning-based computer vision algorithms is the ability to acquire knowledge about the world and use that knowledge to reason about the visual world. Humans can learn about the characteristics of objects and the relationships that occur between them to learn a large variety of visual concepts, often with few examples. This paper investigates the use of structured prior knowledge in the form of knowledge graphs and shows that using this knowledge improves performance on image classification. We build on recent work on end-to-end learning on graphs, introducing the Graph Search Neural Network as a way of efficiently incorporating large knowledge graphs into a vision classification pipeline. We show in a number of experiments that our method outperforms standard neural network baselines for multi-label classification.

**********

Detecting Masked Faces in the Wild With LLE-CNNs

Shiming Ge, Jia Li, Qiting Ye, Zhao Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2682-2690

Detecting masked faces (i.e., faces with occlusions) is a challenging task due to two main reasons: 1)the absence of large datasets of masked faces, and 2)the absence of facial cues from the masked regions. To address these issues, this paper first introduces a dataset with 30,811 Internet images and 35,806 annotated M

Asked FAces, which is denoted as MAFA. Different from many previous datasets, ea ch annotated face in MAFA is partially occluded by mask. By analyzing the charac teristics of masked faces, we propose LLE-CNNs that detect masked face via three major modules. The proposal module first combines two pre-trained CNNs to extra ct candidate facial regions from the input image and represent them with high di mensional descriptors. After that, the embedding module turns such descriptors i nto vectors of weights with respect to the components in pre-trained dictionarie s of representative normal faces and non-faces by using locally linear embedding . In this manner, missing facial cues in the masked regions can be largely recov ered, and the influences of noisy cues introduced by diversified masks can be gr eatly alleviated. Finally, the verification module takes the weight vectors as i nput and identifies real facial regions as well as their accurate positions by j ointly performing the classification and regression tasks within unified CNNs. E xperimental results on MAFA show that the proposed approach significantly outper forms 6 state-of-the-arts by at least 15.6% in detecting masked faces.

**********

UltraStereo: Efficient Learning-Based Matching for Active Stereo Systems

Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir T ankovich, Philip Davidson, Shahram Izadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2691-2700

Efficient estimation of depth from pairs of stereo images is one of the core pro blems in computer vision. We efficiently solve the specialized problem of stereo matching under active illumination using a new learning-based algorithm. This t ype of 'active' stereo i.e. stereo matching where scene texture is augmented by an active light projector is proving compelling for designing depth cameras, lar gely due to improved robustness when compared to time of flight or traditional s tructured light techniques. Our algorithm uses an unsupervised greedy optimizati on scheme that learns features that are discriminative for estimating correspond ences in infrared images. The proposed method optimizes a series of sparse hyper planes that are used at test time to remap all the image patches into a compact binary representation in O(1). The proposed algorithm is cast in a PatchMatch St ereo-like framework, producing depth maps at 500Hz. In contrast to standard stru ctured light methods, our approach generalizes to different scenes, does not req uire tedious per camera calibration procedures and is not adversely affected by interference from overlapping sensors. Extensive evaluations show we surpass the quality and overcome the limitations of current depth sensing technologies.

**********

Learning Features by Watching Objects Move

Deepak Pathak, Ross Girshick, Piotr Dollar, Trevor Darrell, Bharath Hariharan; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 2701-2710

This paper presents a novel yet intuitive approach to unsupervised feature learn ing. Inspired by the human visual system, we explore whether low-level motion-ba sed grouping cues can be used to learn an effective visual representation. Speci fically, we use unsupervised motion-based segmentation on videos to obtain segme nts, which we use as 'pseudo ground truth' to train a convolutional network to s egment objects from a single frame. Given the extensive evidence that motion pla ys a key role in the development of the human visual system, we hope that this s traightforward approach to unsupervised learning will be more effective than cle verly designed 'pretext' tasks studied in the literature. Indeed, our extensive experiments show that this is the case. When used for transfer learning on objec t detection, our representation significantly outperforms previous unsupervised approaches across multiple settings, especially when training data for the targe t task is scarce.

**********

Action-Decision Networks for Visual Tracking With Deep Reinforcement Learning

Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, Jin Young Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2711-2720

This paper proposes a novel tracker which is controlled by sequentially pursuing

actions learned by deep reinforcement learning. In contrast to the existing tra
ckers using deep networks, the proposed tracker is designed to achieve a light c
omputation as well as satisfactory tracking accuracy in both location and scale.
The deep network to control actions is pre-trained using various training seque
nces and fine-tuned during tracking for online adaptation to target and backgrou
nd changes. The pre-training is done by utilizing deep reinforcement learning as
well as supervised learning. The use of reinforcement learning enables even par
tially labeled data to be successfully utilized for semi-supervised learning. Th
rough evaluation of the OTB dataset, the proposed tracker is validated to achiev
e a competitive performance that is three times faster than state-of-the-art, de
ep network-based trackers. The fast version of the proposed method, which operat
es in real-time on GPU, outperforms the state-of-the-art real-time trackers.
**********

Multi-Attention Network for One Shot Learning

Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, Heng Tao
Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2017, pp. 2721-2729

One-shot learning is a challenging problem where the aim is to recognize a class
identified by a single training image. Given the practical importance of one-sh
ot learning, it seems surprising that the rich information present in the class
tag itself has largely been ignored. Most existing approaches restrict the use o
f the class tag to finding similar classes and transferring classifiers or metri
cs learned thereon. We demonstrate here, in contrast, that the class tag can inf
orm one-shot learning as a guide to visual attention on the training image for c
reating the image representation. This is motivated by the fact that human being
s can better interpret a training image if the class tag of the image is underst
ood. Specifically, we design a neural network architecture which takes the seman
tic embedding of the class tag to generate attention maps and uses those attenti
on maps to create the image features for one-shot learning. Note that unlike oth
er applications, our task requires that the learned attention generator can be g
eneralized to novel classes. We show that this can be realized by representing c
lass tags with distributed word embeddings and learning the attention map genera
tor from an auxiliary training set. Also, we design a multiple-attention scheme
to extract richer information from the exemplar image and this leads to substant
ial performance improvement. Through comprehensive experiments, we show that the
proposed approach leads to superior performance over the baseline methods.
**********

G2DeNet: Global Gaussian Distribution Embedding Network and Its Application to V
isual Recognition

Qilong Wang, Peihua Li, Lei Zhang; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2017, pp. 2730-2739

Recently, plugging trainable structural layers into deep convolutional neural ne
tworks (CNNs) as image representations has made promising progress. However, the
re has been little work on inserting parametric probability distributions, which
can effectively model feature statistics, into deep CNNs in an end-to-end manne
r. This paper proposes a Global Gaussian Distribution embedding Network (G2DeNet
) to take a step towards addressing this problem. The core of G2DeNet is a novel
trainable layer of a global Gaussian as an image representation plugged into de
ep CNNs for end-to-end learning. The challenge is that the proposed layer involv
es Gaussian distributions whose space is not a linear space, which makes its for
ward and backward propagations be non-intuitive and non-trivial. To tackle this
issue, we employ a Gaussian embedding strategy which respects the structures of
both Riemannian manifold and smooth group of Gaussians. Based on this strategy,
we construct the proposed global Gaussian embedding layer and decompose it into
two sub-layers: the matrix partition sub-layer decoupling the mean vector and co
variance matrix entangled in the embedding matrix, and the square-rooted, symmet
ric positive definite matrix sub-layer. In this way, we can derive the partial d
erivatives associated with the proposed structural layer and thus allow backprop
agation of gradients. Experimental results on large scale region classification
and fine-grained recognition tasks show that G2DeNet is superior to its counterp

arts, capable of achieving state-of-the-art performance.
**********

Depth From Defocus in the Wild
Huixuan Tang, Scott Cohen, Brian Price, Stephen Schiller, Kiriakos N. Kutulakos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2740-2748
We consider the problem of two-frame depth from defocus in conditions unsuitable for existing methods yet typical of everyday photography: a handheld cellphone camera, a small aperture, a non-stationary scene and sparse surface texture. Our approach combines a global analysis of image content---3D surfaces, deformations, figure-ground relations, textures---with local estimation of joint depth-flow likelihoods in tiny patches. To enable local estimation we (1) derive novel defocus-equalization filters that induce brightness constancy across frames and (2) impose a tight upper bound on defocus blur---just three pixels in radius---through an appropriate choice of the second frame. For global analysis we use a novel piecewise-spline scene representation that can propagate depth and flow across large irregularly-shaped regions. Our experiments show that this combination preserves sharp boundaries and yields good depth and flow maps in the face of significant noise, uncertainty, non-rigidity, and data sparsity.
**********

Fried Binary Embedding for High-Dimensional Visual Features
Weixiang Hong, Junsong Yuan, Sreyasee Das Bhattacharjee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2749-2757
Most existing binary embedding methods prefer compact binary codes (b-dimensional) to avoid high computational and memory cost of projecting high-dimensional visual features (d-dimensional, b < d). We argue that long binary codes (b O(d)) are critical to fully utilize the discriminative power of high-dimensional visual features, and can achieve better results in various tasks such as approximate nearest neighbour search. Generating long binary codes involves large projection matrix and high-dimensional matrix-vector multiplication, thus is memory and compute intensive. To tackle these problems, we propose Fried Binary Embedding (FBE) to decompose the projection matrix using adaptive Fastfood transform, which is the multiplication of several structured matrices. As a result, FBE can reduce the computational complexity from O(d2) to O(dlogd), and memory cost from O(d2) to O(d), respectively. More importantly, by using the structured matrices, FBE can regulate the projection matrix against over-fitting and lead to even better accuracy than using unconstrained projection matrix (like ITQ [4]) with the same long code length. Experimental comparisons with state-of-the-art methods over various visual applications demonstrate both the efficiency and performance advantages of the FBE.
**********

TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering
Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, Gunhee Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2758-2766
Vision and language understanding has emerged as a subject undergoing intense study in Artificial Intelligence. Among many tasks in this line of research, visual question answering (VQA) has been one of the most successful ones, where the goal is to learn a model that understands visual content at region-level details and finds their associations with pairs of questions and answers in the natural language form. Despite the rapid progress in the past few years, most existing work in VQA have focused primarily on images. In this paper, we focus on extending VQA to the video domain and contribute to the literature in three important ways. First, we propose three new tasks designed specifically for video VQA, which require spatio-temporal reasoning from videos to answer questions correctly. Next, we introduce a new large-scale dataset for video VQA named TGIF-QA that extends existing VQA work with our new tasks. Finally, we propose a dual-LSTM based approach with both spatial and temporal attention, and show its effectiveness over conventional VQA techniques through empirical evaluations.

**********
Joint Registration and Representation Learning for Unconstrained Face Identification
Munawar Hayat, Salman H. Khan, Naoufel Werghi, Roland Goecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2767-2776
Recent advances in deep learning have resulted in human-level performances on popular unconstrained face datasets including Labeled Faces in the Wild and YouTube Faces. To further advance research, IJB-A benchmark was recently introduced with more challenges especially in the form of extreme head poses. Registration of such faces is quite demanding and often requires laborious procedures like facial landmark localization. In this paper, we propose a Convolutional Neural Networks based data-driven approach which learns to simultaneously register and represent faces. We validate the proposed scheme on template based unconstrained face identification. Here, a template contains multiple media in the form of images and video frames. Unlike existing methods which synthesize all template media information at feature level, we propose to keep the template media intact. Instead, we represent gallery templates by their trained one-vs-rest discriminative models and then employ a Bayesian strategy which optimally fuses decisions of all medias in a query template. We demonstrate the efficacy of the proposed scheme on IJB-A, YouTube Celebrities and COX datasets where our approach achieves significant relative performance boosts of 3.6%, 21.6% and 12.8% respectively.
**********
Object-Aware Dense Semantic Correspondence
Fan Yang, Xin Li, Hong Cheng, Jianping Li, Leiting Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2777-2785
This work aims to build pixel-to-pixel correspondences between images from the same visual class but with different geometries and visual similarities. This task is particularly challenging because (i) their visual content is similar only on the high-level structure, and (ii) background clutters keep bringing in noises. To address these problems, this paper proposes an object-aware method to estimate per-pixel correspondences from semantic to low-level by learning a classifier for each selected discriminative grid cell and guiding the localization of every pixel under the semantic constraint. Specifically, an Object-aware Hierarchical Graph (OHG) model is constructed to regulate matching consistency from one coarse grid cell containing whole object(s), to fine grid cells covering smaller semantic elements, and finally to every pixel. A guidance layer is introduced as the semantic constraint on local structure matching. In addition, we propose to learn the important high-level structure for each grid cell in an "objectness-driven" way as an alternative to handcrafted descriptors in defining a better visual similarity. The proposed method has been extensively evaluated on various challenging benchmarks and real-world images. The results show that our method significantly outperforms the state-of-the-arts in terms of semantic flow accuracy.
**********
The Misty Three Point Algorithm for Relative Pose
Tobias Palmer, Kalle Astrom, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2786-2794
There is a significant interest in scene reconstruction from underwater images given its utility for oceanic research and for recreational image manipulation. In this paper we propose a novel algorithm for two view camera motion estimation for underwater imagery. Our method leverages the constraints provided by the attenuation properties of water and its effects on the appearance of the color to determine the depth difference of a point with respect to the two observing views of the underwater cameras. Additionally, we propose an algorithm, leveraging the depth differences of three such observed points, to estimate the relative pose of the cameras. Given the unknown underwater attenuation coefficients, our method estimates the relative motion up to scale. The results are represented as a generalized camera. We evaluate our method on both real data and simulated data.
**********

Weakly Supervised Affordance Detection
Johann Sawatzky, Abhilash Srikantha, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2795-2804
Localizing functional regions of objects or affordances is an important aspect of scene understanding and relevant for many robotics applications. In this work, we introduce a pixel-wise annotated affordance dataset of 3090 images containing 9916 object instances. Since parts of an object can have multiple affordances, we address this by a convo- lutional neural network for multilabel affordance segmen- tation. We also propose an approach to train the network from very few keypoint annotations. Our approach achieves a higher affordance detection accuracy than other weakly supervised methods that also rely on keypoint annotations or image annotations as weak supervision.
**********

End-To-End Representation Learning for Correlation Filter Based Tracking
Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2805-2813
The Correlation Filter is an algorithm that trains a linear template to discriminate between images and their translations. It is well suited to object tracking because its formulation in the Fourier domain provides a fast solution, enabling the detector to be re-trained once per frame. Previous works that use the Correlation Filter, however, have adopted features that were either manually designed or trained for a different task. This work is the first to overcome this limitation by interpreting the Correlation Filter learner, which has a closed-form solution, as a differentiable layer in a deep neural network. This enables learning deep features that are tightly coupled to the Correlation Filter. Experiments illustrate that our method has the important practical benefit of allowing lightweight architectures to achieve state-of-the-art performance at high framerates.
**********

Accurate Depth and Normal Maps From Occlusion-Aware Focal Stack Symmetry
Michael Strecke, Anna Alperovich, Bastian Goldluecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2814-2822
We introduce a novel approach to jointly estimate consistent depth and normal maps from 4D light fields, with two main contributions. First, we build a cost volume from focal stack symmetry. However, in contrast to previous approaches, we introduce partial focal stacks in order to be able to robustly deal with occlusions. This idea already yields significanly better disparity maps. Second, even recent sublabel-accurate methods for multi-label optimization recover only a piecewise flat disparity map from the cost volume, with normals pointing mostly towards the image plane. This renders normal maps recovered from these approaches unsuitable for potential subsequent applications. We therefore propose regularization with a novel prior linking depth to normals, and imposing smoothness of the resulting normal field. We then jointly optimize over depth and normals to achieve estimates for both which surpass previous work in accuracy on a recent benchmark.
**********

3D Human Pose Estimation From a Single Image via Distance Matrix Regression
Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2823-2832
This paper addresses the problem of 3D human pose estimation from a single image. We follow a standard two-step pipeline by first detecting the 2D position of the N body joints, and then using these observations to infer 3D pose. For the first step, we use a recent CNN-based detector. For the second step, most existing approaches perform 2N-to-3N regression of the Cartesian joint coordinates. We show that more precise pose estimates can be obtained by representing both the 2D and 3D human poses using NxN distance matrices, and formulating the problem as a 2D-to-3D distance matrix regression. For learning such a regressor we leverage on simple Neural Network architectures, which by construction, enforce positivity and symmetry of the predicted matrices. The approach has also the advantage to naturally handle missing observations and allowing to hypothesize the posi

tion of non-observed joints. Quantitative results on Humaneva and Human3.6M datasets demonstrate consistent performance gains over state-of-the-art. Qualitative evaluation on the images in-the-wild of the LSP dataset, using the regressor learned on Human3.6M, reveals very promising generalization results.
**********

Zero-Shot Action Recognition With Error-Correcting Output Codes
Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, Yunhong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2833-2842
Recently, zero-shot action recognition (ZSAR) has emerged with the explosive growth of action categories. In this paper, we explore ZSAR from a novel perspective by adopting the Error-Correcting Output Codes (dubbed ZSECOC). Our ZSECOC equips the conventional ECOC with the additional capability of ZSAR, by addressing the domain shift problem. In particular, we learn discriminative ZSECOC for seen categories from both category-level semantics and intrinsic data structures. This procedure deals with domain shift implicitly by transferring the well-established correlations among seen categories to unseen ones. Moreover, a simple semantic transfer strategy is developed for explicitly transforming the learned embeddings of seen categories to better fit the underlying structure of unseen categories. As a consequence, our ZSECOC inherits the promising characteristics from ECOC as well as overcomes domain shift, making it more discriminative for ZSAR. We systematically evaluate ZSECOC on three realistic action benchmarks, i.e. Olympic Sports, HMDB51 and UCF101. The experimental results clearly show the superiority of ZSECOC over the state-of-the-art methods.
**********

Multiple Instance Detection Network With Online Instance Classifier Refinement
Peng Tang, Xinggang Wang, Xiang Bai, Wenyu Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2843-2851
Of late, weakly supervised object detection is with great importance in object recognition. Based on deep learning, weakly supervised detectors have achieved many promising results. However, compared with fully supervised detection, it is more challenging to train deep network based detectors in a weakly supervised manner. Here we formulate weakly supervised detection as a Multiple Instance Learning (MIL) problem, where instance classifiers (object detectors) are put into the network as hidden nodes. We propose a novel online instance classifier refinement algorithm to integrate MIL and the instance classifier refinement procedure into a single deep network, and train the network end-to-end with only image-level supervision, i.e., without object location information. More precisely, instance labels inferred from weak supervision are propagated to their spatially overlapped instances to refine instance classifier online. The iterative instance classifier refinement procedure is implemented using multiple streams in deep network, where each stream supervises its latter stream. Weakly supervised object detection experiments are carried out on the challenging PASCAL VOC 2007 and 2012 benchmarks. We obtain 47% mAP on VOC 2007 that significantly outperforms the previous state-of-the-art.
**********

Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild
Shan Li, Weihong Deng, JunPing Du; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2852-2861
Past research on facial expressions have used relatively limited datasets, which makes it unclear whether current methods can be employed in real world. In this paper, we present a novel database, RAF-DB, which contains about 30000 facial images from thousands of individuals. Each image has been individually labeled about 40 times, then EM algorithm was used to filter out unreliable labels. Crowdsourcing reveals that real-world faces often express compound emotions, or even mixture ones. For all we know, RAF-DB is the first database that contains compound expressions in the wild. Our cross-database study shows that the action units of basic emotions in RAF-DB are much more diverse than, or even deviate from, those of lab-controlled ones. To address this problem, we propose a new DLP-CNN (D

eep Locality-Preserving CNN) method, which aims to enhance the discriminative po wer of deep features by preserving the locality closeness while maximizing the i nter-class scatters. The benchmark experiments on the 7-class basic expressions and 11-class compound expressions, as well as the additional experiments on SFEW and CK+ databases, show that the proposed DLP-CNN outperforms the state-of-the-art handcrafted features and deep learning based methods for the expression reco gnition in the wild.
**********

Deep Sketch Hashing: Fast Free-Hand Sketch-Based Image Retrieval
Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, Ling Shao; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2862-2871
Free-hand sketch-based image retrieval (SBIR) is a specific cross-view retrieval task, in which queries are abstract and ambiguous sketches while the retrieval database is formed with natural images. Work in this area mainly focuses on extr acting representative and shared features for sketches and natural images. Howev er, these can neither cope well with the geometric distortion between sketches a nd images nor be feasible for large-scale SBIR due to the heavy continuous-value d distance computation. In this paper, we speed up SBIR by introducing a novel b inary coding method, named Deep Sketch Hashing (DSH), where a semi-heterogeneous deep architecture is proposed and incorporated into an end-to-end binary coding framework. Specifically, three convolutional neural networks are utilized to en code free-hand sketches, natural images and, especially, the auxiliary sketch-to kens which are adopted as bridges to mitigate the sketch-image geometric distort ion. The learned DSH codes can effectively capture the cross-view similarities a s well as the intrinsic semantic correlations between different categories. To t he best of our knowledge, DSH is the first hashing work specifically designed fo r category-level SBIR with an end-to-end deep architecture. The proposed DSH is comprehensively evaluated on two large-scale datasets of TU-Berlin Extension and Sketchy, and the experiments consistently show DSH's superior SBIR accuracies o ver several state-of-the-art methods, while achieving significantly reduced retr ieval time and memory footprint.
**********

Semantic Regularisation for Recurrent Image Annotation
Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, Changyin Sun; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 7, pp. 2872-2880
The "CNN-RNN" design pattern is increasingly widely applied in a variety of imag e annotation tasks including multi-label classification and captioning. Existin g models use the weakly semantic CNN hidden layer or its transform as the image embedding that provides the interface between the CNN and RNN. This leaves the R NN overstretched with two jobs: predicting the visual concepts and modelling the ir correlations for generating structured annotation output. Importantly this ma kes the end-to-end training of the CNN and RNN slow and ineffective due to the d ifficulty of back propagating gradients through the RNN to train the CNN. We pro pose a simple modification to the design pattern that makes learning more effec tive and efficient. Specifically, we propose to use a semantically regularised e mbedding layer as the interface between the CNN and RNN. Regularising the interf ace can partially or completely decouple the learning problems, allowing each to be more effectively trained and jointly training much more efficient. Extensiv e experiments show that state-of-the art performance is achieved on multi-label classification as well as image captioning.
**********

Pyramid Scene Parsing Network
Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 , pp. 2881-2890
Scene parsing is challenging for unrestricted open vocabulary and diverse scenes . In this paper, we exploit the capability of global context information by diff erent-region-based context aggregation through our pyramid pooling module togeth

er with the proposed pyramid scene parsing network (PSPNet). Our global prior re
presentation is effective to produce good quality results on the scene parsing t
ask, while PSPNet provides a superior framework for pixel-level prediction. The
proposed approach achieves state-of-the-art performance on various datasets. It
came first in ImageNet scene parsing challenge 2016, PASCAL VOC 2012 benchmark a
nd Cityscapes benchmark. A single PSPNet yields the new record of mIoU accuracy
85.4% on PASCAL VOC 2012 and accuracy 80.2% on Cityscapes.
**********

On Human Motion Prediction Using Recurrent Neural Networks
Julieta Martinez, Michael J. Black, Javier Romero; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2891-2900
Human motion modelling is a classical problem at the intersection of graphics an
d computer vision, with applications spanning human-computer interaction, motion
 synthesis, and motion prediction for virtual and augmented reality.  Foll
owing the success of deep learning methods in several computer vision tasks, rec
ent work has focused on using deep recurrent neural networks (RNNs) to model hum
an motion, with the goal of learning time-dependent representations that perform
 tasks such as short-term motion prediction and long-term human motion synthesis
. We examine recent work, with a focus on the evaluation methodologies commonly
used in the literature, and show that, surprisingly, state of the art performanc
e can be achieved by a simple baseline that does not attempt to model motion at
all. We investigate this result, and analyze recent RNN methods by looking at th
e architectures, loss functions, and training procedures used in state-of-the-ar
t approaches. We propose three changes to the standard RNN models typically used
 for human motion, which results in a simple and scalable RNN architecture that
obtains state-of-the-art performance on human motion prediction.
**********

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Rea
soning
Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawren
ce Zitnick, Ross Girshick; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2017, pp. 2901-2910
When building artificial intelligence systems that can reason and answer questio
ns about visual data, we need diagnostic tests to analyze our progress and disco
ver short- comings. Existing benchmarks for visual question answer- ing can help
, but have strong biases that models can exploit to correctly answer questions w
ithout reasoning. They also conflate multiple sources of error, making it hard t
o pinpoint model weaknesses. We present a diagnostic dataset that tests a range
of visual reasoning abilities. It contains minimal biases and has detailed annot
ations describing the kind of reasoning each question requires. We use this data
set to analyze a variety of modern visual reasoning systems, providing novel ins
ights into their abilities and limitations.
**********

SST: Single-Stream Temporal Action Proposals
Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, Juan Carlos Niebles
; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
(CVPR), 2017, pp. 2911-2920
Our paper presents a new approach for temporal detection of human actions in lon
g, untrimmed video sequences. We introduce Single-Stream Temporal Action Proposa
ls (SST), a new effective and efficient deep architecture for the generation of
temporal action proposals. Our network can run continuously in a single stream o
ver very long input video sequences, without the need to divide input into short
 overlapping clips or temporal windows for batch processing. We demonstrate empi
rically that our model outperforms the state-of-the-art on the task of temporal
action proposal generation, while achieving some of the fastest processing speed
s in the literature. Finally, we demonstrate that using SST proposals in conjunc
tion with existing action classifiers results in improved state-of-the-art tempo
ral action detection performance.
**********

Kernel Pooling for Convolutional Neural Networks

Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2921-2930

Convolutional Neural Networks (CNNs) with Bilinear Pooling, initially in their full form and later using compact representations, have yielded impressive performance gains on a wide range of visual tasks, including fine-grained visual categorization, visual question answering, face recognition, and description of texture and style. The key to their success lies in the spatially invariant modeling of pairwise (2nd order) feature interactions. In this work, we propose a general pooling framework that captures higher order interactions of features in the form of kernels. We demonstrate how to approximate kernels such as Gaussian RBF up to a given order using compact explicit feature maps in a parameter-free manner. Combined with CNNs, the composition of the kernel can be learned from data in an end-to-end fashion via error back-propagation. The proposed kernel pooling scheme is evaluated in terms of both kernel approximation error and visual recognition accuracy. Experimental evaluations demonstrate state-of-the-art performance on commonly used fine-grained recognition datasets.

**********

Subspace Clustering via Variance Regularized Ridge Regression

Chong Peng, Zhao Kang, Qiang Cheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2931-2940

Spectral clustering based subspace clustering methods have emerged recently. When the inputs are 2-dimensional (2D) data, most existing clustering methods convert such data to vectors as preprocessing, which severely damages spatial information of the data. In this paper, we propose a novel subspace clustering method for 2D data with enhanced capability of retaining spatial information for clustering. It seeks two projection matrices and simultaneously constructs a linear representation of the projected data, such that the sought projections help construct the most expressive representation with the most variational information. We regularize our method based on covariance matrices directly obtained from 2D data, which have much smaller size and are more computationally amiable. Moreover, to exploit nonlinear structures of the data, a nonlinear version is proposed, which constructs an adaptive manifold according to updated projections. The learning processes of projections, representation, and manifold thus mutually enhance each other, leading to a powerful data representation. Efficient optimization procedures are proposed, which generate non-increasing objective value sequence with theoretical convergence guarantee. Extensive experimental results confirm the effectiveness of proposed method.

**********

Efficient Global Point Cloud Alignment Using Bayesian Nonparametric Mixtures

Julian Straub, Trevor Campbell, Jonathan P. How, John W. Fisher III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2941-2950

Point cloud alignment is a common problem in computer vision and robotics, with applications ranging from 3D object recognition to reconstruction. We propose a novel approach to the alignment problem that utilizes Bayesian nonparametrics to describe the point cloud and surface normal densities, and branch and bound (BB) optimization to recover the relative transformation. BB uses a novel, refinable, near-uniform tessellation of rotation space using 4D tetrahedra, leading to more efficient optimization compared to the common axis-angle tessellation. We provide objective function bounds for pruning given the proposed tessellation, and prove that BB converges to the optimum of the cost function along with providing its computational complexity. Finally, we empirically demonstrate the efficiency of the proposed approach as well as its robustness to real-world conditions such as missing data and partial overlap.

**********

The Incremental Multiresolution Matrix Factorization Algorithm

Vamsi K. Ithapu, Risi Kondor, Sterling C. Johnson, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2951-2960

Multiresolution analysis and matrix factorization are foundational tools in comp
uter vision. In this work, we study the interface between these two distinct top
ics and obtain techniques to uncover hierarchical block structure in symmetric m
atrices -- an important aspect in the success of many vision problems. Our new a
lgorithm, the incremental multiresolution matrix factorization, uncovers such st
ructure one feature at a time, and hence scales well to large matrices. We descr
ibe how this multiscale analysis goes much farther than what a direct "global" f
actorization of the data can identify.  We evaluate the efficacy of the resultin
g factorizations for relative leveraging within regression tasks using medical i
maging data.  We also use the factorization on representations learned by popul
ar deep networks,  providing evidence of their ability to infer semantic relatio
nships even when they are not explicitly trained to do so. We show that this alg
orithm can be used as an exploratory tool to improve the network architecture, a
nd within numerous other settings in vision.
**********

CATS: A Color and Thermal Stereo Benchmark
Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Ne
al, Brian Phelan, Kelly Sherbondy, Chandra Kambhamettu; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2961-296
9
Stereo matching is a well researched area using visible-band color cameras. Ther
mal images are typically lower resolution, have less texture, and are noisier co
mpared to their visible-band counterparts and are more challenging for stereo ma
tching algorithms. Previous benchmarks for stereo matching either focus entirely
 on visible-band cameras or contain only a single thermal camera. We present the
 Color And Thermal Stereo (CATS) benchmark, a dataset consisting of stereo therm
al, stereo color, and cross-modality image pairs with high accuracy ground truth
 (< 2mm) generated from a LiDAR. We scanned 100 cluttered indoor and 80 outdoor
scenes featuring challenging environments and conditions. CATS contains approxim
ately 1400 images of pedestrians, vehicles, electronics, and other thermally int
eresting objects in different environmental conditions, including nighttime, day
time, and foggy scenes. Ground truth was projected to each of the four cameras t
o generate color-color, thermal-thermal, and cross-modality disparity maps. We d
evelop a semi-automatic LiDAR to camera alignment procedure that does not requir
e a calibration target. We compare state-of-the-art algorithms to baseline the d
ataset and show that in the thermal and cross modalities there is still much roo
m for improvement. We expect our dataset to provide researchers with a more dive
rse set of imaged locations, objects, and modalities than previous benchmarks fo
r stereo matching.
**********

Deep Image Matting
Ning Xu, Brian Price, Scott Cohen, Thomas Huang; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2970-2979
Image matting is a fundamental computer vision problem and has many applications
. Previous algorithms have poor performance when an image has similar foreground
 and background colors or complicated textures. The main reasons are prior metho
ds 1) only use low-level features and 2) lack high-level context. In this paper,
 we propose a novel deep learning based algorithm that can tackle both these pro
blems. Our deep model has two parts. The first part is a deep convolutional enco
der-decoder network that takes an image and the corresponding trimap as inputs a
nd predict the alpha matte of the image. The second part is a small convolutiona
l network that refines the alpha matte predictions of the first network to have
more accurate alpha values and sharper edges. In addition, we also create a larg
e-scale image matting dataset including 49300 training images and 1000 testing i
mages. We evaluate our algorithm on the image matting benchmark, our testing set
, and a wide variety of real images. Experimental results clearly demonstrate th
e superiority of our algorithm over previous methods.
**********

The Surfacing of Multiview 3D Drawings via Lofting and Occlusion Reasoning
Anil Usumezbas, Ricardo Fabbri, Benjamin B. Kimia; Proceedings of the IEEE Confe

rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2980-2989
The three-dimensional reconstruction of scenes from multiple views has made impr essive strides in recent years, chiefly by methods correlating isolated feature points, intensities, or curvilinear structure. In the general setting, i.e., wit hout requiring controlled acquisition, limited number of objects, abundant patte rns on objects, or object curves to follow particular models, the majority of th ese methods produce unorganized point clouds, meshes, or voxel representations o f the reconstructed scene, with some exceptions producing 3D drawings as network s of curves. Many applications, e.g., robotics, urban planning, industrial desig n, and hard surface modeling, however, require structured representations which make explicit 3D curves, surfaces, and their spatial relationships. Reconstructi ng surface representations can now be constrained by the 3D drawing acting like a scaffold to hang on the computed representations, leading to increased robustn ess and quality of reconstruction. This paper presents one way of completing suc h 3D drawings with surface reconstructions, by exploring occlusion reasoning thr ough lofting algorithms.
**********

One-Shot Metric Learning for Person Re-Identification
Slawomir Bak, Peter Carr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2990-2999
Re-identification of people in surveillance footage must cope with drastic varia tions in color, background, viewing angle and a person's pose. Supervised techn iques are often the most effective, but require extensive annotation which is in feasible for large camera networks. Unlike previous supervised learning approac hes that require hundreds of annotated subjects, we learn a metric using a nove l one-shot learning approach. We first learn a deep texture representation from intensity images with Convolutional Neural Networks (CNNs). When training a CNN using only intensity images, the learned embedding is color-invariant and shows high performance even on unseen datasets without fine-tuning. To account for di fferences in camera color distributions, we learn a color metric using a single pair of ColorChecker images. The proposed one-shot learning achieves performance that is competitive with supervised methods, but uses only a single example rat her than the hundreds required for the fully supervised case. Compared with sem i-supervised and unsupervised state-of-the-art methods, our approach yields sign ificantly higher accuracy.
**********

Richer Convolutional Features for Edge Detection
Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, Xiang Bai; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3000-3009
In this paper, we propose an accurate edge detector using richer convolutional f eatures (RCF). Since objects in natural images possess various scales and aspect ratios, learning the rich hierarchical representations is very critical for edg e detection. CNNs have been proved to be effective for this task. In addition, t he convolutional features in CNNs gradually become coarser with the increase of the receptive fields. According to these observations, we attempt to adopt riche r convolutional features in such a challenging vision task. The proposed network fully exploits multiscale and multilevel information of objects to perform the image-to-image prediction by combining all the meaningful convolutional features in a holistic manner. Using VGG16 network, we achieve state-of-the-art performa nce on several available datasets. When evaluating on the well-known BSDS500 ben chmark, we achieve ODS F-measure of 0.811 while retaining a fast speed (8 FPS). Besides, our fast version of RCF achieves ODS F-measure of 0.806 with 30 FPS.
**********

Video Segmentation via Multiple Granularity Analysis
Rui Yang, Bingbing Ni, Chao Ma, Yi Xu, Xiaokang Yang; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3010-3019
We introduce a Multiple Granularity Analysis framework for video segmentation in a coarse-to-fine manner. We cast video segmentation as a spatio-temporal superp ixel labeling problem. Benefited from the bounding volume provided by off-the-sh

elf object trackers, we estimate the foreground/ background super-pixel labeling using the spatiotemporal multiple instance learning algorithm to obtain coarse foreground/background separation within the volume. We further refine the segmentation mask in the pixel level using the graph-cut model. Extensive experiments on benchmark video datasets demonstrate the superior performance of the proposed video segmentation algorithm.

**********

Learning Cross-Modal Embeddings for Cooking Recipes and Food Images

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3020-3028

In this paper, we introduce Recipe1M, a new large-scale, structured corpus of over 1m cooking recipes and 800k food images. As the largest publicly available collection of recipe data, Recipe1M affords the ability to train high-capacity models on aligned, multi-modal data. Accordingly, we train a neural network to find a joint embedding of recipes and images that yields impressive results on an image-recipe retrieval task. Additionally, we demonstrate that regularization via the addition of a high-level, semantic classification objective improves performance to rival that of humans and enables semantic vector arithmetic. We postulate that these embeddings will provide a basis for further exploration of the Recipe1M dataset and food and cooking in general.

**********

Locality-Sensitive Deconvolution Networks With Gated Fusion for RGB-D Indoor Semantic Segmentation

Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, Kaiqi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3029-3037

This paper focuses on indoor semantic segmentation using RGB-D data. Although the commonly used deconvolution networks (DeconvNet) have achieved impressive results on this task, we find there is still room for improvements in two aspects. One is about the boundary segmentation. DeconvNet aggregates large context to predict the label of each pixel, inherently limiting the segmentation precision of object boundaries. The other is about RGB-D fusion. Recent state-of-the-art methods generally fuse RGB and depth networks with equal-weight score fusion, regardless of the varying contributions of the two modalities on delineating different categories in different scenes. To address the two problems, we first propose a locality-sensitive DeconvNet (LS-DeconvNet) to refine the boundary segmentation over each modality. LS-DeconvNet incorporates locally visual and geometric cues from the raw RGB-D data into each DeconvNet, which is able to learn to upsample the coarse convolutional maps with large context whilst recovering sharp object boundaries. Towards RGB-D fusion, we introduce a gated fusion layer to effectively combine the two LS-DeconvNets. This layer can learn to adjust the contributions of RGB and depth over each pixel for high-performance object recognition. Experiments on the large-scale SUN RGB-D dataset and the popular NYU-Depth v2 dataset show that our approach achieves new state-of-the-art results for RGB-D indoor semantic segmentation.

**********

One-To-Many Network for Visually Pleasing Compression Artifacts Reduction

Jun Guo, Hongyang Chao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3038-3047

We consider the compression artifacts reduction problem, where a compressed image is transformed into an artifact-free image. Recent approaches for this problem typically train a one-to-one mapping using a per-pixel L_2 loss between the outputs and the ground-truths. We point out that these approaches used to produce overly smooth results, and PSNR doesn't reflect their real performance. In this paper, we propose a one-to-many network, which measures output quality using a perceptual loss, a naturalness loss, and a JPEG loss. We also avoid grid-like artifacts during deconvolution using a "shift-and-average" strategy. Extensive experimental results demonstrate the dramatic visual improvement of our approach over the state of the arts.

**********
Recurrent Modeling of Interaction Context for Collective Activity Recognition
Minsi Wang, Bingbing Ni, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3048-3056
Modeling of high order interactional context, e.g., group interaction, lies in the central of collective/group activity recognition. However, most of the previous activity recognition methods do not offer a flexible and scalable scheme to handle the high order context modeling problem. To explicitly address this fundamental bottleneck, we propose a recurrent interactional context modeling scheme based on LSTM network. By utilizing the information propagation/aggregation capability of LSTM, the proposed scheme unifies the interactional feature modeling process for single person dynamics, intra-group (e.g., persons within a group) and inter-group(e.g., group to group)interactions. The proposed high order context modeling scheme produces more discriminative/descriptive interactional features. It is very flexible to handle a varying number of input instances (e.g., different number of persons in a group or different number of groups) and linearly scalable to high order context modeling problem. Extensive experiments on two benchmark collective/group activity datasets demonstrate the effectiveness of the proposed method.
**********
Hierarchical Multimodal Metric Learning for Multimodal Classification
Heng Zhang, Vishal M. Patel, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3057-3065
Multimodal classification arises in many computer vision tasks such as object classification and image retrieval. The idea is to utilize multiple sources (modalities) measuring the same instance to improve the overall performance compared to using a single source (modality). The varying characteristics exhibited by multiple modalities make it necessary to simultaneously learn the corresponding metrics. In this paper, we propose a multiple metrics learning algorithm for multimodal data. Metric of each modality is a product of two matrices: one matrix is modality specific, the other is enforced to be shared by all the modalities. The learned metrics can improve multimodal classification accuracy and experimental results on four datasets show that the proposed algorithm outperforms existing learning algorithms based on multiple metrics as well as other approaches tested on these datasets. Specifically, we report 95.0% object instance recognition accuracy, 89.2% object category recognition accuracy on the multi-view RGB-D dataset and 52.3% scene category recognition accuracy on SUN RGB-D dataset.
**********
Probabilistic Temporal Subspace Clustering
Behnam Gholami, Vladimir Pavlovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3066-3075
Subspace clustering is a common modeling paradigm used to identify constituent modes of variation in data with locally linear structure. These structures are common to many problems in computer vision, including modeling time series of complex human motion. However classical subspace clustering algorithms learn the relationships within a set of data without considering the temporal dependency and then use a separate clustering step (e.g., spectral clustering) for final segmentation. Moreover, these, frequently optimization-based, algorithms assume that all observations have complete features. In contrast in real-world applications, some features are often missing, which results in incomplete data and substantial performance degeneration of these approaches. In this paper, we propose a unified non-parametric generative framework for temporal subspace clustering to segment data drawn from a sequentially ordered union of subspaces that deals with the missing features in a principled way. The non-parametric nature of our generative model makes it possible to infer the number of subspaces and their dimension automatically from data. Experimental results on human action datasets demonstrate that the proposed model consistently outperforms other state-of-the-art subspace clustering approaches.
**********
Detecting Visual Relationships With Deep Relational Networks

Bo Dai, Yuqi Zhang, Dahua Lin; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2017, pp. 3076-3086
Relationships among objects play a crucial role in image understanding. Despite
 the great success of deep learning techniques in recognizing individual objects,
 reasoning about the relationships among objects remains a challenging task. Pre
vious methods often treat this as a classification problem, considering each typ
e of relationship (e.g. "ride") or each distinct visual phrase (e.g. "person- ri
de-horse") as a category. Such approaches are faced with significant difficultie
s caused by the high diversity of visual appearance for each kind of relationshi
ps or the large number of distinct visual phrases. We propose an integrated fram
ework to tackle this problem. At the heart of this framework is the Deep Relatio
nal Network, a novel formulation designed specifically for exploiting the statis
tical dependencies between objects and their relationships. On two large data se
ts, the proposed method achieves substantial improvement over state-of-the-art.
**********

Discover and Learn New Objects From Documentaries
Kai Chen, Hang Song, Chen Change Loy, Dahua Lin; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3087-3096
Despite the remarkable progress in recent years, detecting objects in a new cont
ext remains a challenging task. Detectors learned from a public dataset can only
 work with a fixed list of categories, while training from scratch usually requi
res a large amount of training data with detailed annotations. This work aims to
 explore a novel approach -- learning object detectors from documentary films in
 a weakly supervised manner. This is inspired by the observation that documentar
ies often provide dedicated exposition of certain object categories, where visua
l presentations are aligned with subtitles. We believe that object detectors can
 be learned from such a rich source of information. Towards this goal, we develo
p a joint probabilistic framework, where individual pieces of information, inclu
ding video frames and subtitles, are brought together via both visual and lingui
stic links. On top of this formulation, we further derive a weakly supervised le
arning algorithm, where object model learning and training set mining are unifie
d in an optimization procedure. Experimental results on a real world dataset dem
onstrate that this is an effective approach to learning new object detectors.
**********

Spatio-Temporal Vector of Locally Max Pooled Features for Action Recognition in
Videos
Ionut Cosmin Duta, Bogdan Ionescu, Kiyoharu Aizawa, Nicu Sebe; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3
097-3106
We introduce Spatio-Temporal Vector of Locally Max Pooled Features (ST-VLMPF), a
 super vector-based encoding method specifically designed for local deep feature
s encoding.  The proposed method addresses an important problem of video underst
anding: how to build a video representation that incorporates the CNN features o
ver the entire video. Feature assignment is carried out at two levels, by using
the similarity and spatio-temporal information. For each assignment we build a s
pecific encoding, focused on the nature of deep features,  with the goal to capt
ure the highest feature responses from the highest neuron activation of the netw
ork. Our ST-VLMPF clearly provides a more reliable video representation than  so
me of the most widely used and powerful encoding approaches (Improved Fisher Vec
tors and Vector of Locally  Aggregated Descriptors), while maintaining a low com
putational complexity. We conduct experiments on three action recognition datase
ts: HMDB51, UCF50 and UCF101. Our pipeline obtains  state-of-the-art results.
**********

Specular Highlight Removal in Facial Images
Chen Li, Stephen Lin, Kun Zhou, Katsushi Ikeuchi; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3107-3116
We present a method for removing specular highlight reflections in facial images
 that may contain varying illumination colors. This is accurately achieved throu
gh the use of physical and statistical properties of human skin and faces. We em
ploy a melanin and hemoglobin based model to represent the diffuse color variati

ons in facial skin, and utilize this model to constrain the highlight removal so lution in a manner that is effective even for partially saturated pixels. The re moval of highlights is further facilitated through estimation of directionally v ariant illumination colors over the face, which is done while taking advantage o f a statistically-based approximation of facial geometry. An important practical feature of the proposed method is that the skin color model is utilized in a wa y that does not require color calibration of the camera. Moreover, this approach does not require assumptions commonly needed in previous highlight removal tech niques, such as uniform illumination color or piecewise-constant surface colors. We validate this technique through comparisons to existing methods for removing specular highlights.
**********

Radiometric Calibration From Faces in Images
Chen Li, Stephen Lin, Kun Zhou, Katsushi Ikeuchi; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3117-3126
We present a method for radiometric calibration of cameras from a single image t hat contains a human face. This technique takes advantage of a low-rank property that exists among certain skin albedo gradients because of the pigments within the skin. This property becomes distorted in images that are captured with a non -linear camera response function, and we perform radiometric calibration by solv ing for the inverse response function that best restores this low-rank property in an image. Although this work makes use of the color properties of skin pigmen ts, we show that this calibration is unaffected by the color of scene illuminati on or the sensitivities of the camera's color filters. Our experiments validate this approach on a variety of images containing human faces, and show that faces can provide an important source of calibration data in images where existing ra diometric calibration techniques perform poorly.
**********

What Can Help Pedestrian Detection?
Jiayuan Mao, Tete Xiao, Yuning Jiang, Zhimin Cao; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3127-3136
Aggregating extra features has been considered as an effective approach to boost traditional pedestrian detection methods. However, there is still a lack of stu dies on whether and how CNN-based pedestrian detectors can benefit from these ex tra features. The first contribution of this paper is exploring this issue by ag gregating extra features into CNN-based pedestrian detection framework. Through extensive experiments, we evaluate the effects of different kinds of extra featu res quantitatively. Moreover, we propose a novel network architecture, namely Hy perLearner, to jointly learn pedestrian detection as well as the given extra fea ture. By multi-task training, HyperLearner is able to utilize the information of given features and improve detection performance without extra inputs in infere nce. The experimental results on multiple pedestrian benchmarks validate the eff ectiveness of the proposed HyperLearner.
**********

StyleNet: Generating Attractive Visual Captions With Styles
Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, Li Deng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3137-31 46
We propose a novel framework named StyleNet to address the task of generating at tractive captions for images and videos with different styles. To this end, we devise a novel model component, named factored LSTM, which automatically distill s the style factors in the monolingual text corpus. Then at runtime, we can expl icitly control the style in the caption generation process so as to produce attr active visual captions with the desired style. Our approach achieves this goal b y leveraging two sets of data: 1) factual image/video-caption paired data, and 2 ) stylized monolingual text data (e.g., romantic and humorous sentences). We sho w experimentally that StyleNet outperforms existing approaches for generating vi sual captions with different styles, measured in both automatic and human evalua tion metrics on the newly collected FlickrStyle10K image caption dataset, which contains 10K Flickr images with corresponding humorous and romantic captions.

**********
Image Super-Resolution via Deep Recursive Residual Network
Ying Tai, Jian Yang, Xiaoming Liu; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 3147-3155
Recently, Convolutional Neural Network (CNN) based models have achieved great su ccess in Single Image Super-Resolution (SISR). Owing to the strength of deep net works, these CNN models learn an effective nonlinear mapping from the low-resolu tion input image to the high-resolution target image, at the cost of requiring e normous parameters. This paper proposes a very deep CNN model (up to 52 convolut ional layers) named Deep Recursive Residual Network (DRRN) that strives for deep yet concise networks. Specifically, residual learning is adopted, both in globa l and local manners, to mitigate the difficulty of training very deep networks; recursive learning is used to control the model parameters while increasing the depth. Extensive benchmark evaluation shows that DRRN significantly outperforms state of the art in SISR, while utilizing far fewer parameters. Code is availabl e at https://github.com/tyshiwo/DRRN_CVPR17.
**********
Residual Attention Network for Image Classification
Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaoga ng Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3156-3164
In this work, we propose "Residual Attention Network", a convolutional neural ne twork using attention mechanism which can incorporate with state-of-art feed for ward network architecture in an end-to-end training fashion. Our Residual Attent ion Network is built by stacking Attention Modules which generate attention-awar e features. The attention-aware features from different modules change adaptivel y as layers going deeper. Inside each Attention Module, bottom-up top-down feedf orward structure is used to unfold the feedforward and feedback attention proces s into a single feedforward process. Importantly, we propose attention residual learning to train very deep Residual Attention Networks which can be easily scal ed up to hundreds of layers. Extensive analyses are conducted on CIFAR-10 and CI FAR-100 datasets to verify the effectiveness of every module mentioned above. Ou r Residual Attention Network achieves state-of-the-art object recognition perfor mance on three benchmark datasets including CIFAR-10 (3.90% error), CIFAR-100 (2 0.45% error) and ImageNet (4.8% single model and single crop, top-5 error). Note that, our method achieves 0.6% top-1 accuracy improvement with 46% trunk depth and 69% forward FLOPs comparing to ResNet-200. The experiment also demonstrates that our network is robust against noisy labels.
**********
End-To-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering
Youngjae Yu, Hyungjin Ko, Jongwook Choi, Gunhee Kim; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3165-3173
We propose a high-level concept word detector that can be integrated with any vi deo-to-language models. It takes a video as input and generates a list of concep t words as useful semantic priors for language generation models. The proposed w ord detector has two important properties. First, it does not require any extern al knowledge sources for training. Second, the proposed word detector is trainab le in an end-to-end manner jointly with any video-to-language models. To effecti vely exploit the detected words, we also develop a semantic attention mechanism that selectively focuses on the detected concept words and fuse them with the wo rd encoding and decoding in the language model. In order to demonstrate that the proposed approach indeed improves the performance of multiple video-to-language tasks, we participate in all the four tasks of LSMDC 2016. Our approach has won three of them, including fill-in-the-blank, multiple-choice test, and movie ret rieval.
**********
Semantic Autoencoder for Zero-Shot Learning
Elyor Kodirov, Tao Xiang, Shaogang Gong; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2017, pp. 3174-3183

Existing zero-shot learning (ZSL) models typically learn a projection function f
rom a feature space to a semantic embedding space (e.g. attribute space). Howeve
r, such a projection function is only concerned with predicting the training see
n class  semantic representation (e.g. attribute prediction) or classification.
When applied to test data, which in the context of ZSL contains different (unsee
n) classes without training data, a ZSL model typically suffers from the project
 domain shift problem. In this work, we present a novel solution to ZSL based on
 learning a Semantic AutoEncoder (SAE). Taking the encoder-decoder paradigm, an
encoder aims to project a visual feature vector into the semantic space as in th
e existing ZSL models. However, the decoder exerts an additional constraint, tha
t is, the projection/code must be able to reconstruct the original visual featur
e. We show that with this additional reconstruction constraint, the learned  pro
jection function from the seen classes is able to generalise better to the new u
nseen classes. Importantly, the encoder and decoder are linear and symmetric whi
ch enable us to develop an extremely efficient learning algorithm. Extensive exp
eriments on six benchmark datasets demonstrate that the proposed SAE outperforms
 significantly the existing ZSL models with the additional benefit of lower comp
utational cost. Furthermore, when the SAE is applied to supervised clustering pr
oblem, it also beats the state-of-the-art.
**********

Co-Occurrence Filter
Roy J. Jevnisek, Shai Avidan; Proceedings of the IEEE Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2017, pp. 3184-3192
Co-occurrence Filter (CoF) is a boundary preserving filter. It is based on the B
ilateral Filter (BF) but instead of using a Gaussian on the range values to pres
erve edges it relies on a co-occurrence matrix. Pixel values that co-occur frequ
ently in the image (i.e., inside textured regions) will have a high weight in th
e co-occurrence matrix. This, in turn, means  that such pixel pairs will be aver
aged and hence smoothed, regardless of their intensity differences. On the other
 hand, pixel values that rarely co-occur (i.e., across texture boundaries) will
have a low weight in the co-occurrence matrix. As a result, they will not be ave
raged and the boundary between them will be preserved. The CoF therefore extends
 the  BF to deal with boundaries, not just edges. It learns co-occurrences direc
tly from the image. We can achieve various filtering results by directing it to
learn the co-occurrence matrix from a part of the image, or a different image. W
e give the definition of the filter, discuss how to use it with color images and
 show several use cases.
**********

Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer
Cascade
Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, Xiaoou Tang; Proceedings of t
he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.
3193-3202
We propose a novel deep layer cascade (LC) method to improve the accuracy and sp
eed of semantic segmentation. Unlike the conventional model cascade (MC) that is
 composed of multiple independent models, LC treats a single deep model as a cas
cade of several sub-models. Earlier sub-models are trained to handle easy and co
nfident regions, and they progressively feed-forward harder regions to the next
sub-model for processing. Convolutions are only calculated on these regions to r
educe computations. The proposed method possesses several advantages. First, LC
classifies most of the easy regions in the shallow stage and makes deeper stage
focuses on a few hard regions. Such an adaptive and 'difficulty-aware' learning
improves segmentation performance. Second, LC accelerates both training and test
ing of deep network thanks to early decisions in the shallow stage. Third, in co
mparison to MC, LC is an end-to-end trainable framework, allowing joint learning
 of all sub-models. We evaluate our method on PASCAL VOC and Cityscapes datasets
, achieving state-of-the-art performance and fast speed.
**********

Deeply Supervised Salient Object Detection With Short Connections
Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, Philip H. S. Torr

; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3203-3212
Recent progress on saliency detection is substantial, benefiting mostly from the explosive development of Convolutional Neural Networks (CNNs). Semantic segmentation and saliency detection algorithms developed lately have been mostly based on Fully Convolutional Neural Networks (FCNs). There is still a large room for improvement over the generic FCN models that do not explicitly deal with the scale-space problem. Holisitcally-Nested Edge Detector (HED) provides a skip-layer structure with deep supervision for edge and boundary detection, but the performance gain of HED on saliency detection is not obvious. In this paper, we propose a new saliency method by introducing short connections to the skip-layer structures within the HED architecture. Our framework provides rich multi-scale feature maps at each layer, a property that is critically needed to perform segment detection. Our method produces state-of-the-art results on 5 widely tested salient object detection benchmarks, with advantages in terms of efficiency (0.08 seconds per image), effectiveness, and simplicity over the existing algorithms.
**********

CityPersons: A Diverse Dataset for Pedestrian Detection
Shanshan Zhang, Rodrigo Benenson, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3213-3221
Convnets have enabled significant progress in pedestrian detection recently, but there are still open questions regard- ing suitable architectures and training data. We revisit CNN design and point out key adaptations, enabling plain Fas- terRCNN to obtain state-of-the-art results on the Caltech dataset. To achieve further improvement from more and better data, we introduce CityPersons, a new set of person annotations on top of the Cityscapes dataset. The di- versity of CityPersons allows us for the first time to train one single CNN model that generalizes well over mul- tiple benchmarks. Moreover, with additional training with CityPersons, we obtain top results using FasterRCNN on Caltech, improving especially for more difficult cases (heavy occlusion and small scale) and providing higher loc- alization quality.
**********

Generalized Rank Pooling for Activity Recognition
Anoop Cherian, Basura Fernando, Mehrtash Harandi, Stephen Gould; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3222-3231
Most popular deep models for action recognition split video sequences into short sub-sequences consisting of a few frames; frame-based features are then pooled for recognizing the activity. Usually, this pooling step discards the temporal order of the frames, which could otherwise be used for better recognition. Towards this end, we propose a novel pooling method, generalized rank pooling (GRP), that takes as input, features from the intermediate layers of a CNN that is trained on tiny sub-sequences, and produces as output the parameters of a subspace which (i) provides a low-rank approximation to the features and (ii) preserves their temporal order. We propose to use these parameters as a compact representation for the video sequence, which is then used in a classification setup. We formulate an objective for computing this subspace as a Riemannian optimization problem on the Grassmann manifold, and propose an efficient conjugate gradient scheme for solving it. Experiments on several activity recognition datasets show that our scheme leads to state-of-the-art performance.
**********

Deep Cross-Modal Hashing
Qing-Yuan Jiang, Wu-Jun Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3232-3240
Due to its low storage cost and fast query speed, cross-modal hashing (CMH) has been widely used for similarity search in multimedia retrieval applications. However, most existing CMH methods are based on hand-crafted features which might not be optimally compatible with the hash-code learning procedure. As a result, existing CMH methods with hand-crafted features may not achieve satisfactory performance. In this paper, we propose a novel CMH method, called deep cross-modal h

ashing (DCMH), by integrating feature learning and hash-code learning intothe sa me framework. DCMH is an end-to-end learning framework with deep neural networks , one for each modality, to perform feature learning from scratch. Experiments o n three real datasets with image-text modalities show that DCMH can outperform o ther baselines to achieve the state-of-the-art performance in cross-modal retrie val applications.
**********

Revisiting Metric Learning for SPD Matrix Based Visual Representation
Luping Zhou, Lei Wang, Jianjia Zhang, Yinghuan Shi, Yang Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 32 41-3249
The success of many visual recognition tasks largely depends on a good similarit y measure, and distance metric learning plays an important role in this regard. Meanwhile, Symmetric Positive Definite (SPD) matrix is receiving increased atten tion for feature representation in multiple computer vision applications. Howeve r, distance metric learning on SPD matrices has not been sufficiently researched . A few existing works approached this by learning either d^2 xp or d xk transfo rmation matrix for dxd SPD matrices. Different from these methods, this paper pr oposes a new member to the family of distance metric learning for SPD matrices. It learns only d parameters to adjust the eigenvalues of the SPD matrices throug h an efficient optimisation scheme. Also, it is shown that the proposed method c an be interpreted as learning a sample-specific transformation matrix, instead o f the fixed transformation matrix learned for all the samples in the existing wo rks. The optimised d parameters can be used to "massage" the SPD matrices for be tter discrimination while still keeping them in the original space. From this pe rspective, the proposed method complements, rather than competes with, the exist ing linear-transformation-based methods, as the latter can always be applied to the output of the former to perform distance metric learning in further. The pro posed method has been tested on multiple SPD-based visual representation data se ts used in the literature, and the results demonstrate its interesting propertie s and attractive performance.
**********

CNN-Based Patch Matching for Optical Flow With Thresholded Hinge Embedding Loss
Christian Bailer, Kiran Varanasi, Didier Stricker; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3250-3259
Learning based approaches have not yet achieved their full potential in optical flow estimation, where their performance still trails heuristic approaches. In t his paper, we present a CNN based patch matching approach for optical flow estim ation. An important contribution of our approach is a novel thresholded loss for Siamese networks. We demonstrate that our loss performs clearly better than exi sting losses. It also allows to speed up training by a factor of 2 in our tests. Furthermore, we present a novel way for calculating CNN based features for diff erent image scales, which performs better than existing methods. We also discuss new ways of evaluating the robustness of trained features for the application o f patch matching for optical flow. An interesting discovery in our paper is that low-pass filtering of feature maps can increase the robustness of features crea ted by CNNs. We proved the competitive performance of our approach by submitting it to the KITTI 2012, KITTI 2015 and MPI-Sintel evaluation portals where we obt ained state-of-the-art results on all three datasets.
**********

A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Video s
Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konra d Schindler, Marc Pollefeys, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3260-3269
Motivated by the limitations of existing multi-view stereo benchmarks, we presen t a novel dataset for this task. Towards this goal, we recorded a variety of ind oor and outdoor scenes using a high-precision laser scanner and captured both hi gh-resolution DSLR imagery as well as synchronized low-resolution stereo videos with varying fields-of-view. To align the images with the laser scans, we propos

e a robust technique which minimizes photometric errors conditioned on the geome try. In contrast to previous datasets, our benchmark provides novel challenges a nd covers a diverse set of viewpoints and scene types, ranging from natural scen es to man-made indoor and outdoor environments. Furthermore, we provide data at significantly higher temporal and spatial resolution. Our benchmark is the first to cover the important use case of hand-held mobile devices while also providin g high-resolution DSLR camera images. We make our datasets and an online evaluat ion server available at http://www.eth3d.net.
**********

Snapshot Hyperspectral Light Field Imaging
Zhiwei Xiong, Lizhi Wang, Huiqun Li, Dong Liu, Feng Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3270-327 8
This paper presents the first snapshot hyperspectral light field imager in pract ice. Specifically, we design a novel hybrid camera system to obtain two compleme ntary measurements that sample the angular and spectral dimensions respectively. To recover the full 5D hyperspectral light field from the severely undersampled measurements, we then propose an efficient computational reconstruction algorit hm by exploiting the large correlations across the angular and spectral dimensio ns through self-learned dictionaries. Simulation on an elaborate hyperspectral l ight field dataset validates the effectiveness of the proposed approach. Hardwar e experimental results demonstrate that, for the first time to our knowledge, a 5D hyperspectral light field containing 9x9 angular views and 27 spectral bands can be acquired in a single shot.
**********

Zero-Shot Recognition Using Dual Visual-Semantic Mapping Paths
Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, Yueting Zhuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3279-3287
Zero-shot recognition aims to accurately recognize objects of unseen classes by using a shared visual-semantic mapping between the image feature space and the s emantic embedding space. This mapping is learned on training data of seen class es and is expected to have transfer ability to unseen classes. In this paper, we tackle this problem by exploiting the intrinsic relationship between the semant ic space manifold and the transfer ability of visual-semantic mapping. We formal ize their connection and cast zero-shot recognition as a joint optimization prob lem. Motivated by this, we propose a novel framework for zero-shot recognition, which contains dual visual-semantic mapping paths. Our analysis shows this frame work can not only apply prior semantic knowledge to infer underlying semantic ma nifold in the image feature space, but also generate optimized semantic embeddin g space, which can enhance the transfer ability of the visual-semantic mapping t o unseen classes. The proposed method is evaluated for zero-shot recognition on four benchmark datasets, achieving outstanding results.
**********

A New Representation of Skeleton Sequences for 3D Action Recognition
Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3288-3297
This paper presents a new method for 3D action recognition with skeleton sequenc es (i.e., 3D trajectories of human skeleton joints). The proposed method first t ransforms each skeleton sequence into three clips each consisting of several fra mes for spatial temporal feature learning using deep neural networks. Each clip is generated from one channel of the cylindrical coordinates of the skeleton seq uence. Each frame of the generated clips represents the temporal information of the entire skeleton sequence, and incorporates one particular spatial relationsh ip between the joints. The entire clips include multiple frames with different s patial relationships, which provide useful spatial structural information of the human skeleton. We propose to use deep convolutional neural networks to learn l ong-term temporal information of the skeleton sequence from the frames of the ge nerated clips, and then use a Multi-Task Learning Network (MTLN) to jointly proc

ess all frames of the clips in parallel to incorporate spatial structural inform ation for action recognition. Experimental results clearly show the effectivenes s of the proposed new representation and feature learning method for 3D action r ecognition.
**********

Efficient Linear Programming for Dense CRFs
Thalaiyasingam Ajanthan, Alban Desmaison, Rudy Bunel, Mathieu Salzmann, Philip H . S. Torr, M. Pawan Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3298-3306
The fully connected conditional random field (CRF) with Gaussian pairwise potent ials has proven popular and effective for multi-class semantic segmentation. Whi le the energy of a dense CRF can be minimized accurately using a linear programm ing (LP) relaxation, the state-of-the-art algorithm is too slow to be useful in practice. To alleviate this deficiency, we introduce an efficient LP minimizatio n algorithm for dense CRFs. To this end, we develop a proximal minimization fram ework, where the dual of each proximal problem is optimized via block coordinate descent. We show that each block of variables can be efficiently optimized. Spe cifically, for one block, the problem decomposes into significantly smaller subp roblems, each of which is defined over a single pixel. For the other block, the problem is optimized via conditional gradient descent. This has two advantages: 1) the conditional gradient can be computed in a time linear in the number of pi xels and labels; and 2) the optimal step size can be computed analytically. Our experiments on standard datasets provide compelling evidence that our approach o utperforms all existing baselines including the previous LP based approach for d ense CRFs.
**********

Learning Deep Match Kernels for Image-Set Classification
Haoliang Sun, Xiantong Zhen, Yuanjie Zheng, Gongping Yang, Yilong Yin, Shuo Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C VPR), 2017, pp. 3307-3316
Image-set classification has recently generated great popularity due to its wide spread applications in computer vision. The great challenges arise from effectiv ely and efficiently measuring the similarity between image sets with high inter- class ambiguity and huge intra-class variability. In this paper, we propose deep match kernels (DMK) to directly measure the similarity between image sets in th e match kernel framework. Specifically, we build deep local match kernels betwee n images upon arc-cosine kernels, which can faithfully characterize the similari ty between images by mimicking deep neural networks; we introduce anchors to agg regate those deep local match kernels into a global match kernel between image s ets, which is learned in a supervised way by kernel alignment and therefore more discriminative. The DMK provides the first match kernel framework for image-set classification, which removes specific assumptions usually required in previous approaches and is computationally more efficient. We conduct extensive experime nts on four datasets for three diverse image-set classification tasks. The DMK a chieves high performance and consistently surpasses state-of-the-art methods, sh owing its great effectiveness for image-set classification.
**********

A Deep Regression Architecture With Two-Stage Re-Initialization for High Perform ance Facial Landmark Detection
Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, Xi Zhou; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3317-3326
Regression based facial landmark detection methods usually learns a series of re gression functions to update the landmark positions from an initial estimation. Most of existing approaches focus on learning effective mapping functions with r obust image features to improve performance. The approach to dealing with the in itialization issue, however, receives relatively fewer attentions. In this paper , we present a deep regression architecture with two-stage re-initialization to explicitly deal with the initialization problem. At the global stage, given an i mage with a rough face detection result, the full face region is firstly re-init

ialized by a supervised spatial transformer network to a canonical shape state a
nd then trained to regress a coarse landmark estimation. At the local stage, dif
ferent face parts are further separately re-initialized to their own canonical s
hape states, followed by another regression subnetwork to get the final estimati
on. Our proposed deep architecture is trained from end to end and obtains promis
ing results using different kinds of unstable initialization. It also achieves s
uperior performances over many competing algorithms.
**********

Product Manifold Filter: Non-Rigid Shape Correspondence via Kernel Density Estim
ation in the Product Space
Matthias Vestner, Roee Litman, Emanuele Rodola, Alex Bronstein, Daniel Cremers;
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C
VPR), 2017, pp. 3327-3336
Many algorithms for the computation of correspondences between deformable shapes
 rely on some variant of nearest neighbor matching in a descriptor space. Such a
re, for example, various point-wise correspondence recovery algorithms used as a
 post-processing stage in the functional correspondence framework. Such frequent
ly used techniques implicitly make restrictive assumptions (e.g., nearisometry)
on the considered shapes and in practice suffer from lack of accuracy and result
 in poor surjectivity. We propose an alternative recovery technique capable of g
uaranteeing a bijective correspondence and producing significantly higher accura
cy and smoothness. Unlike other methods our approach does not depend on the assu
mption that the analyzed shapes are isometric. We derive the proposed method fro
m the statistical framework of kernel density estimation and demonstrate its per
formance on several challenging deformable 3D shape matching datasets.
**********

Elastic Shape-From-Template With Spatially Sparse Deforming Forces
Abed Malti, Cedric Herzet; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2017, pp. 3337-3345
Current Elastic SfT (Shape from Template) methods are based on l2-norm minimizat
ion. None can accurately recover the spatial location of the acting forces since
 l2-norm based  minimization tends to find the best tradeoff among noisy data to
 fit an elastic model. In this work, we study shapes that are deformed with spat
ially sparse set of forces. We propose two formulations for a new class of SfT p
roblems dubbed here SLE-SfT (Sparse Linear Elastic-SfT). The First ideal formula
tion uses an l0-norm to minimize the cardinal of non-zero components of the defo
rming forces. The second relaxed formulation uses an l1-norm to minimize the sum
 of absolute values of force components. These new formulations do not use Solid
 Boundary Constraints (SBC) which are usually needed to rigidly position the sha
pe in the frame of the deformed image. We introduce the Projective Elastic Space
 Property (PESP) that jointly encodes the reprojection constraint and the elasti
c model. We prove that filling this property is necessary and sufficient for the
 relaxed formulation to: (i) retrieve the ground-truth 3D deformed shape, (ii) r
ecover the right spatial domain of non-zero deforming forces. (iii) It also prov
es that we can rigidly place the deformed shape in the image frame without using
 SBC. Finally, we prove that when filling PESP, resolving the relaxed formulatio
n provides the same ground-truth solution as the ideal formulation. Results with
 simulated and real data show substantial improvements in recovering the deforme
d shapes as well as the spatial location of the deforming forces.
**********

A General Framework for Curve and Surface Comparison and Registration With Orien
ted Varifolds
Irene Kaltenmark, Benjamin Charlier, Nicolas Charon; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3346-3355
This paper introduces a general setting for the construction of data fidelity me
trics between oriented or non-oriented geometric shapes like curves, curve sets
or surfaces. These metrics are based on the representation of shapes as distribu
tions of their local tangent or normal vectors and the definition of reproducing
 kernels on these spaces. The construction, that combines in one common setting
and extends the previous frameworks of currents and varifolds, provides a very l

arge class of kernel metrics which can be easily computed without requiring any kind of parametrization of shapes and which are smooth enough to give robustness to certain imperfections that could result e.g. from bad segmentation. We then give a sense, with synthetic examples, of the versatility and potentialities of such metrics when used in various problems like shape comparison, clustering and diffeomorphic registration.
**********

BranchOut: Regularization for Online Ensemble Tracking With Convolutional Neural Networks
Bohyung Han, Jack Sim, Hartwig Adam; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2017, pp. 3356-3365
We propose an extremely simple but effective regularization technique of convolu tional neural networks (CNNs), referred to as BranchOut, for online ensemble tra cking. Our algorithm employs a CNN for target representation, which has a commo n convolutional layers but has multiple branches of fully connected layers. For better regularization, a subset of branches in the CNN are selected randomly fo r online learning whenever target appearance models need to be updated. Each br anch may have a different number of layers to maintain variable abstraction leve ls of target appearances. BranchOut with multi-level target representation allow s us to learn robust target appearance models with diversity and handle various challenges in visual tracking problem effectively. The proposed algorithm is ev aluated in standard tracking benchmarks and shows the state-of-the-art performan ce even without additional pretraining on external tracking sequences.
**********

Expert Gate: Lifelong Learning With a Network of Experts
Rahaf Aljundi, Punarjay Chakravarty, Tinne Tuytelaars; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3366-3375
In this paper we introduce a model of lifelong learning, based on a Network of E xperts. New tasks / experts are learned and added to the model sequentially, bui lding on what was learned before. To ensure scalability of this process, data fr om previous tasks cannot be stored and hence is not available when learning a ne w task. A critical issue in such context, not addressed in the literature so far , relates to the decision which expert to deploy at test time. We introduce a s et of gating autoencoders that learn a representation for the task at hand, and, at test time, automatically forward the test sample to the relevant expert. Thi s also brings memory efficiency as only one expert network has to be loaded into memory at any given time. Further, the autoencoders inherently capture the rela tedness of one task to another, based on which the most relevant prior model to be used for training a new expert, with fine-tuning or learning-without-forgetti ng, can be selected. We evaluate our method on image classification and video pr ediction problems.
**********

AdaScan: Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human A ction Recognition in Videos
Amlan Kar, Nishant Rai, Karan Sikka, Gaurav Sharma; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3376-3385
We propose a novel method for temporally pooling frames in a video for the task of human action recognition. The method is motivated by the observation that the re are only a small number of frames which, together, contain sufficient informa tion to discriminate an action class present in a video, from the rest. The prop osed method learns to pool such discriminative and informative frames, while dis carding a majority of the non-informative frames in a single temporal scan of th e video. Our algorithm does so by continuously predicting the discriminative imp ortance of each video frame and subsequently pooling them in a deep learning fra mework. We show the effectiveness of our proposed pooling method on standard ben chmarks where it consistently improves on baseline pooling methods, with both RG B and optical flow based Convolutional networks. Further, in combination with co mplementary video representations, we show results that are competitive with res pect to the state-of-the-art results on two challenging and publicly available b enchmark datasets.

**********
Learning Motion Patterns in Videos
Pavel Tokmakov, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3386-3394
The problem of determining whether an object is in motion, irrespective of camera motion, is far from being solved. We address this challenging task by learning motion patterns in videos. The core of our approach is a fully convolutional network, which is learned entirely from synthetic video sequences, and their ground-truth optical flow and motion segmentation. This encoder-decoder style architecture first learns a coarse representation of the optical flow field features, and then refines it iteratively to produce motion labels at the original high-resolution. We further improve this labeling with an objectness map and a conditional random field, to account for errors in optical flow, and also to focus on moving "things" rather than "stuff". The output label of each pixel denotes whether it has undergone independent motion, i.e., irrespective of camera motion. We demonstrate the benefits of this learning framework on the moving object segmentation task, where the goal is to segment all objects in motion. Our approach outperforms the top method on the recently released DAVIS benchmark dataset, comprising real-world sequences, by 5.6%. We also evaluate on the Berkeley motion segmentation database, achieving state-of-the-art results.
**********
Provable Self-Representation Based Outlier Detection in a Union of Subspaces
Chong You, Daniel P. Robinson, Rene Vidal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3395-3404
Many computer vision tasks involve processing large amounts of data contaminated by outliers, which need to be detected and rejected. While outlier detection methods based on robust statistics have existed for decades, only recently have methods based on sparse and low-rank representation been developed along with guarantees of correct outlier detection when the inliers lie in one or more low-dimensional subspaces. This paper proposes a new outlier detection method that combines tools from sparse representation with random walks on a graph. By exploiting the property that data points can be expressed as sparse linear combinations of each other, we obtain an asymmetric affinity matrix among data points, which we use to construct a weighted directed graph. By defining a suitable Markov Chain from this graph, we establish a connection between inliers/outliers and essential/inessential states of the Markov chain, which allows us to detect outliers by using random walks. We provide a theoretical analysis that justifies the correctness of our method under geometric and connectivity assumptions. Experimental results on image databases demonstrate its superiority with respect to state-of-the-art sparse and low-rank outlier detection methods.
**********
Deep Structured Learning for Facial Action Unit Intensity Estimation
Robert Walecki, Ognjen (Oggi) Rudovic, Vladimir Pavlovic, Bjoern Schuller, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3405-3414
We consider the task of automated estimation of facial expression intensity. This involves estimation of multiple output variables (facial action units --- AUs) that are structurally dependent. Their structure arises from statistically induced co-occurrence patterns of AU intensity levels. Modeling this structure is critical for improving the estimation performance; however, this performance is bounded by the quality of the input features extracted from face images. The goal of this paper is to model these structures and estimate complex feature representations simultaneously by combining conditional random field (CRF) encoded AU dependencies with deep learning. To this end, we propose a novel Copula CNN deep learning approach for modeling multivariate ordinal variables. Our model accounts for ordinal structure in output variables and their non-linear dependencies via copula functions modeled as cliques of a CRF. These are jointly optimized with deep CNN feature encoding layers using a newly introduced balanced batch iterative training algorithm. We demonstrate the effectiveness of our approach on the task of AU intensity estimation on two benchmark datasets. We show that joint lea

rning of the deep features and the target output structure results in significan t performance gains compared to existing structured deep models and deep models for analysis of facial expressions.
**********

Joint Detection and Identification Feature Learning for Person Search
Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 341 5-3424
Existing person re-identification benchmarks and methods mainly focus on matchin g cropped pedestrian images between queries and candidates. However, it is diffe rent from real-world scenarios where the annotations of pedestrian bounding boxe s are unavailable and the target person needs to be searched from a gallery of w hole scene images. To close the gap, we propose a new deep learning framework fo r person search. Instead of breaking it down into two separate tasks---pedestria n detection and person re-identification, we jointly handle both aspects in a si ngle convolutional neural network. An Online Instance Matching (OIM) loss functi on is proposed to train the network effectively, which is scalable to datasets w ith numerous identities. To validate our approach, we collect and annotate a lar ge-scale benchmark dataset for person search. It contains 18,184 images, 8,432 i dentities, and 96,143 pedestrian bounding boxes. Experiments show that our frame work outperforms other separate approaches, and the proposed OIM loss function c onverges much faster and better than the conventional Softmax loss.
**********

Learning to Align Semantic Segmentation and 2.5D Maps for Geolocalization
Anil Armagan, Martin Hirzer, Peter M. Roth, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 342 5-3432
We present an efficient method for geolocalization in urban environments startin g from a coarse estimate of the location provided by a GPS and using a simple un textured 2.5D model of the surrounding buildings. Our key contribution is a nove l efficient and robust method to optimize the pose: We train a Deep Network to p redict the best direction to improve a pose estimate, given a semantic segmentat ion of the input image and a rendering of the buildings from this estimate. We t hen iteratively apply this CNN until converging to a good pose. This approach av oids the use of reference images of the surroundings, which are difficult to acq uire and match, while 2.5D models are broadly available. We can therefore apply it to places unseen during training.
**********

LCR-Net: Localization-Classification-Regression for Human Pose
Gregory Rogez, Philippe Weinzaepfel, Cordelia Schmid; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3433-3441
We propose an end-to-end architecture for joint 2D and 3D human pose estimation in natural images. Key to our approach is the generation and scoring of a number of pose proposals per image, which allows us to predict 2D and 3D pose of multi ple people simultaneously. Hence, our approach does not require an approximate l ocalization of the humans for initialization. Our architecture, named LCR-Net, c ontains 3 main components: 1) the pose proposal generator that suggests potentia l poses at different locations in the image; 2) a classifier that scores the dif ferent pose proposals; and 3) a regressor that refines pose proposals both in 2D and 3D. All three stages share the convolutional feature layers and are trained jointly. The final pose estimation is obtained by integrating over neighboring pose hypotheses, which is shown to improve over a standard non maximum suppressi on algorithm. Our approach significantly outperforms the state of the art in 3D pose estimation on Human3.6M, a controlled environment. Moreover, it shows promi sing results on real images for both single and multi-person subsets of the MPII 2D pose benchmark.
**********

Primary Object Segmentation in Videos Based on Region Augmentation and Reduction
Yeong Jun Koh, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2017, pp. 3442-3450

A novel algorithm to segment a primary object in a video sequence is proposed in this work. First, we generate candidate regions for the primary object using both color and motion edges. Second,we estimate initial primary object regions, by exploiting the recurrence property of the primary object. Third, we augment the initial regions with missing parts or reducing them by excluding noisy parts repeatedly. This augmentation and reduction process (ARP) identifies the primary object region in each frame. Experimental results demonstrate that the proposed algorithm significantly outperforms the state-of-the-art conventional algorithms on recent benchmark datasets.

**********

Deep 360 Pilot: Learning a Deep Agent for Piloting Through 360deg Sports Videos

Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, Min Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3451-3460

Watching a 360* sports video requires a viewer to continuously select a viewing angle, either through a sequence of mouse clicks or head movements. To relieve the viewer from this "360 piloting" task, we propose "deep 360 pilot" - a deep learning-based agent for piloting through 360* sports videos automatically. At each frame, the agent observes a panoramic image and has the knowledge of previously selected viewing angles. The task of the agent is to shift the current viewing angle (i.e. action) to the next preferred one (i.e., goal). We propose to directly learn an online policy of the agent from data. Specifically, we leverage a state-of-the-art object detector to propose a few candidate objects of interest (yellow boxes in Fig. 1). Then, a recurrent neural network is used to select the main object (green dash boxes in Fig. 1). Given the main object and previously selected viewing angles, our method regresses a shift in viewing angle to move to the next one. We use the policy gradient technique to jointly train our pipeline, by minimizing: (1) a regression loss measuring the distance between the selected and ground truth viewing angles, (2) a smoothness loss encouraging smooth transition in viewing angle, and (3) maximizing an expected reward of focusing on a foreground object. To evaluate our method, we built a new 360-Sports video dataset consisting of five sports domains. We trained domain-specific agents and achieved the best performance on viewing angle selection accuracy and users' preference compared to [54] and other baselines.

**********

Learning and Refining of Privileged Information-Based RNNs for Action Recognition From Depth Sequences

Zhiyuan Shi, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3461-3470

Existing RNN-based approaches for action recognition from depth sequences require either skeleton joints or hand-crafted depth features as inputs. An end-to-end manner, mapping from raw depth maps to action classes, is non-trivial to design due to the fact that: 1) single channel map lacks texture thus weakens the discriminative power; 2) relatively small set of depth training data. To address these challenges, we propose to learn an RNN driven by privileged information (PI) in three-steps: An encoder is pre-trained to learn a joint embedding of depth appearance and PI (i.e. skeleton joints). The learned embedding layers are then tuned in the learning step, aiming to optimize the network by exploiting PI in a form of multi-task loss. However, exploiting PI as a secondary task provides little help to improve the performance of a primary task (i.e. classification) due to the gap between them. Finally, a bridging matrix is defined to connect two tasks by discovering latent PI in the refining step. Our PI-based classification loss maintains a consistency between latent PI and predicted distribution. The latent PI and network are iteratively estimated and updated in an expectation-maximization procedure. The proposed learning process provides greater discriminative power to model subtle depth difference, while helping avoid overfitting the scarcer training data. Our experiments show significant performance gains over state-of-the-art methods on three public benchmark datasets and our newly collected Blanket dataset.

**********

Simultaneous Facial Landmark Detection, Pose and Deformation Estimation Under Facial Occlusion
Yue Wu, Chao Gou, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3471-3480
Facial landmark detection, head pose estimation, and facial deformation analysis are typical facial behavior analysis tasks in computer vision. The existing methods usually perform each task independently and sequentially, ignoring their interactions. To tackle this problem, we propose a unified framework for simultaneous facial landmark detection, head pose estimation, and facial deformation analysis, and the proposed model is robust to facial occlusion. Following a cascade procedure augmented with model-based head pose estimation, we iteratively update the facial landmark locations, facial occlusion, head pose and facial deformation until convergence. The experimental results on benchmark databases demonstrate the effectiveness of the proposed method for simultaneous facial landmark detection, head pose and facial deformation estimation, even if the images are under facial occlusion.
**********
A Domain Based Approach to Social Relation Recognition
Qianru Sun, Bernt Schiele, Mario Fritz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3481-3490
Social relations are the foundation of human daily life. Developing techniques to analyze such relations from visual data bears great potential to build machines that better understand us and are capable of interacting with us at a social level. Previous investigations have remained partial due to the overwhelming diversity and complexity of the topic and consequently have only focused on a handful of social relations. In this paper, we argue that the domain-based theory from social psychology is a great starting point to systematically approach this problem. The theory provides coverage of all aspects of social relations and equally is concrete and predictive about the visual attributes and behaviors defining the relations included in each domain. We provide the first dataset built on this holistic conceptualization of social life that is composed of a hierarchical label space of social domains and social relations. We also contribute the first models to recognize such domains and relations and find superior performance for attribute based features. Beyond the encouraging performance of the attribute based approach, we also find interpretable features that are in accordance with the predictions from social psychology literature. Beyond our findings, we believe that our contributions more tightly interleave visual recognition and social psychology theory that has the potential to complement the theoretical work in the area with empirical and data-driven models of social life.
**********
Fractal Dimension Invariant Filtering and Its CNN-Based Implementation
Hongteng Xu, Junchi Yan, Nils Persson, Weiyao Lin, Hongyuan Zha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3491-3499
Fractal analysis has been widely used in computer vision, especially in texture image processing and texture analysis.  The key concept of fractal-based image model is the fractal dimension, which is invariant to bi-Lipschitz transformation of image, and thus capable of representing intrinsic structural information of image robustly.  However, the invariance of fractal dimension generally does not hold after filtering, which limits the application of fractal-based image model.  In this paper, we propose a novel fractal dimension invariant filtering (FDIF) method, extending the invariance of fractal dimension to filtering operations.  Utilizing the notion of local self-similarity, we first develop a local fractal model for images. By adding a nonlinear post-processing step behind anisotropic filter banks, we demonstrate that the proposed filtering method is capable of preserving the local invariance of the fractal dimension of image. Meanwhile, we show that the FDIF method can be re-instantiated approximately via a CNN-based architecture, where the convolution layer extracts anisotropic structure of image and the nonlinear layer enhances the structure via preserving local fractal dimension of image.  The proposed filtering method provides us with a novel geomet

ric interpretation of CNN-based image model. Focusing on a challenging image pr
ocessing task --- detecting complicated curves from the texture-like images, the
proposed method obtains superior results to the state-of-art approaches.
**********

Transformation-Grounded Image Generation Network for Novel 3D View Synthesis
Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, Alexander C. Berg; Proceed
ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2
017, pp. 3500-3509
We present a transformation-grounded image generation network for novel 3D view
synthesis from a single image. Our approach first explicitly infers the parts of
the geometry visible both in the input and novel views and then casts the remai
ning synthesis problem as image completion. Specifically, we both predict a flow
to move the pixels from the input to the novel view along with a novel visibili
ty map that helps deal with occulsion/disocculsion. Next, conditioned on those i
ntermediate results, we hallucinate (infer) parts of the object invisible in the
input image. In addition to the new network structure, training with a combinat
ion of adversarial and perceptual loss results in a reduction in common artifact
s of novel view synthesis such as distortions and holes, while successfully gene
rating high frequency details and preserving visual aspects of the input image.
We evaluate our approach on a wide range of synthetic and real examples. Both qu
alitative and quantitative results show our method achieves significantly better
results compared to existing methods.
**********

Noise-Blind Image Deblurring
Meiguang Jin, Stefan Roth, Paolo Favaro; Proceedings of the IEEE Conference on C
omputer Vision and Pattern Recognition (CVPR), 2017, pp. 3510-3518
We present a novel approach to noise-blind deblurring, the problem of deblurring
an image with known blur, but unknown noise level. We introduce an efficient an
d robust solution based on a Bayesian framework using a smooth generalization of
the 0-1 loss. A novel bound allows the calculation of very high-dimensional int
egrals in closed form. It avoids the degeneracy of Maximum a-Posteriori (MAP) es
timates and leads to an effective noise-adaptive scheme. Moreover, we drasticall
y accelerate our algorithm by using Majorization Minimization (MM) without intro
ducing any approximation or boundary artifacts. We further speed up convergence
by turning our algorithm into a neural network termed GradNet, which is highly p
arallelizable and can be efficiently trained. We demonstrate that our noise-blin
d formulation can be integrated with different priors and significantly improves
existing deblurring algorithms in the noise-blind and in the known-noise case.
Furthermore, GradNet leads to state-of-the-art performance across different nois
e levels, while retaining high computational efficiency.
**********

Multi-Scale FCN With Cascaded Instance Aware Segmentation for Arbitrary Oriented
Word Spotting in the Wild
Dafang He, Xiao Yang, Chen Liang, Zihan Zhou, Alexander G. Ororbi II, Daniel Kif
er, C. Lee Giles; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2017, pp. 3519-3528
Scene text detection has attracted great attention these years. Text potentiall
y exist in a wide variety of images or videos and play an important role in unde
rstanding the scene. In this paper, we present a novel text detection algorithm
which is composed of two cascaded steps: (1) a multi-scale fully convolutional
neural network (FCN) is proposed to extract text block regions; (2) a novel in
stance (word or line) aware segmentation is designed to further remove false pos
itives and obtain word instances. The proposed algorithm can accurately localiz
e word or text line in arbitrary orientations, including curved text lines which
cannot be handled in a lot of other frameworks. Our algorithm achieved state-o
f-the-art performance in ICDAR 2013 (IC13), ICDAR 2015 (IC15) and CUTE80 and Str
eet View Text (SVT) benchmark datasets.
**********

Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image S
egmentation

Anirban Roy, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2017, pp. 3529-3538
This paper addresses the problem of weakly supervised semantic image segmentatio
n. Our goal is to label every pixel in a new image, given only image-level objec
t labels associated with training images. Our problem statement differs from com
mon semantic segmentation, where pixel-wise annotations are typically assumed av
ailable in training. We specify a novel deep architecture which fuses three dist
inct computation processes toward semantic segmentation -- namely, (i) the botto
m-up computation of neural activations in a CNN for the image-level prediction o
f object classes; (ii) the top-down estimation of conditional likelihoods of the
 CNN's activations given the predicted objects, resulting in probabilistic atten
tion maps per object class; and (iii) the lateral attention-message passing from
 neighboring neurons at the same CNN layer. The fusion of (i)-(iii) is realized
via a conditional random field as recurrent network aimed at generating a smooth
 and boundary-preserving segmentation. Unlike existing work, we formulate a unif
ied end-to-end learning of all components of our deep architecture. Evaluation o
n the benchmark PASCAL VOC 2012 dataset demonstrates that we outperform reasonab
le weakly supervised baselines and state-of-the-art approaches.
**********

Multiple People Tracking by Lifted Multicut and Person Re-Identification
Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, Bernt Schiele; Proceedings of the
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 353
9-3548
Tracking multiple persons in a monocular video of a crowded scene is a challengi
ng task. Humans can master it even if they loose track of a person locally by re
-identifying the same person based on their appearance. Care must be taken acros
s long distances, as similar-looking persons need not be identical. In this work
, we propose a novel graph-based formulation that links and clusters person hypo
theses over time by solving an instance of a minimum cost lifted multicut proble
m. Our model generalizes previous works by introducing a mechanism for adding lo
ng-range attractive connections between nodes in the graph without modifying the
 original set of feasible solutions. This allows us to reward tracks that assign
 detections of similar appearance to the same person in a way that does not intr
oduce implausible solutions. To effectively match hypotheses over longer tempora
l gaps we develop new deep architectures for re-identification of people. They c
ombine holistic representations extracted with deep networks and body pose layou
t obtained with a state-of-the-art pose estimation model. We demonstrate the eff
ectiveness of our formulation by reporting a new state-of-the-art for the MOT16
benchmark.
**********

Filter Flow Made Practical: Massively Parallel and Lock-Free
Sathya N. Ravi, Yunyang Xiong, Lopamudra Mukherjee, Vikas Singh; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.
 3549-3558
This paper is inspired by a relatively recent work of Seitz and Baker which intr
oduced the so-called Filter Flow model. Filter flow finds the transformation rel
ating a pair of (or multiple) images by identifying a large set of local linear
filters; imposing additional constraints on certain structural properties of the
se filters enables Filter Flow to serve as a general "one stop" construction for
 a spectrum of problems in vision: from optical flow to defocus to stereo to aff
ine alignment. The idea is beautiful yet the benefits are not borne out in pract
ice because of significant computational challenges. This issue makes most (if n
ot all) deployments for practical vision problems out of reach. The key thrust o
f our work is to identify mathematically (near) equivalent reformulations of thi
s model that can eliminate this serious limitation. We demonstrate via a detaile
d optimization-focused development that Filter Flow can indeed be solved fairly
efficiently for a wide range of instantiations. We derive efficient algorithms,
perform extensive theoretical analysis focused on convergence and parallelizatio
n and show how results competitive with the state of the art for many applicatio
ns can be achieved with negligible application specific adjustments or post-proc

essing. The actual numerical scheme is easy to understand and, implement (30 lin
es in Matlab) -- this development will enable Filter Flow to be a viable general
 solver and testbed for numerous applications in the community, going forward.
**********

Acquiring Axially-Symmetric Transparent Objects Using Single-View Transmission I
maging
Jaewon Kim, Ilya Reshetouski, Abhijeet Ghosh; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3559-3567
We propose a novel, practical solution for high quality reconstruction of axiall
y-symmetric transparent objects. While a special case, such transparent objects
are ubiquitous in the real world. Common examples of these are glasses, goblets,
 tumblers, carafes, etc., that can have very unique and visually appealing forms
 making their reconstruction interesting for vision and graphics applications. O
ur acquisition setup involves imaging such objects from a single viewpoint while
 illuminating them from directly behind with a few patterns emitted by an LCD pa
nel. Our reconstruction step is then based on optimization of the object's geome
try and its refractive index to minimize the difference between observed and sim
ulated transmission/refraction of rays passing through the object. We exploit th
e object's axial symmetry as a strong shape prior which allows us to achieve rob
ust reconstruction from a single viewpoint using a simple, commodity acquisition
 setup. We demonstrate high quality reconstruction of several common rotationall
y symmetric as well as more complex n-fold symmetric transparent objects with ou
r approach.
**********

Online Graph Completion: Multivariate Signal Recovery in Computer Vision
Won Hwa Kim, Mona Jalal, Seongjae Hwang, Sterling C. Johnson, Vikas Singh; Proce
edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
 2017, pp. 3568-3576
The adoption of "human-in-the-loop" paradigms in computer vision and machine lea
rning is leading to various applications where the actual data acquisition (e.g.
, human supervision) and the underlying inference algorithms are closely interwi
ned. While classical work in active learning provides effective solutions when t
he learning module involves classification and regression tasks, many practical
issues such as partially observed measurements, financial constraints and even a
dditional distributional or structural aspects of the data typically fall outsid
e the scope of this treatment. For instance, with sequential acquisition of part
ial measurements of data that manifest as a matrix (or tensor), novel strategies
 for completion (or collaborative filtering) of the remaining entries have only
been studied recently. Motivated by vision problems where we seek to annotate a
large dataset of images via a crowdsourced platform or alternatively, complement
 results from a state-of-the-art object detector using human feedback, we study
the "completion" problem defined on graphs, where requests for additional measur
ements must be made sequentially. We design the optimization model in the Fourie
r domain of the graph describing how ideas based on adaptive submodularity provi
de algorithms that work well in practice. On a large set of images collected fro
m Imgur, we see promising results on images that are otherwise difficult to cate
gorize. We also show applications to an experimental design problem in neuroimag
ing.
**********

OctNet: Learning Deep 3D Representations at High Resolutions
Gernot Riegler, Ali Osman Ulusoy, Andreas Geiger; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3577-3586
We present OctNet, a representation for deep learning with sparse 3D data. In co
ntrast to existing models, our representation enables 3D convolutional networks
which are both deep and high resolution. Towards this goal, we exploit the spars
ity in the input data to hierarchically partition the space using a set of unbal
anced octrees where each leaf node stores a pooled feature representation. This
allows to focus memory allocation and computation to the relevant dense regions
and enables deeper networks without compromising resolution. We demonstrate the
utility of our OctNet representation by analyzing the impact of resolution on se

veral 3D tasks including 3D object classification, orientation estimation and po
int cloud labeling.
**********

Non-Local Color Image Denoising With Convolutional Neural Networks
Stamatios Lefkimmiatis; Proceedings of the IEEE Conference on Computer Vision an
d Pattern Recognition (CVPR), 2017, pp. 3587-3596
We propose a novel deep network architecture for grayscale and color image denoi
sing that is based on a non-local image model. Our motivation for the overall de
sign of the proposed network stems from variational methods that exploit the i
nherent non-local self-similarity property of natural images. We build on this
concept and introduce deep networks that perform non-local processing and at th
e same time they significantly benefit from discriminative learning. Experiments
on the Berkeley segmentation dataset, comparing several state-of-the-art method
s, show that the proposed non-local models achieve the best reported denoising p
erformance both for grayscale and color images for all the tested noise levels.
It is also worth noting that this increase in performance comes at no extra cost
on the capacity of the network compared to existing alternative deep network ar
chitectures. In addition, we highlight a direct link of the proposed non-local m
odels to convolutional neural networks. This connection is of significant import
ance since it allows our models to take full advantage of the latest advances on
GPU computing in deep learning and makes them amenable to efficient implementat
ions through their inherent parallelism.
**********

Slow Flow: Exploiting High-Speed Cameras for Accurate and Diverse Optical Flow R
eference Data
Joel Janai, Fatma Guney, Jonas Wulff, Michael J. Black, Andreas Geiger; Proceedi
ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20
17, pp. 3597-3607
Existing optical flow datasets are limited in size and variability due to the di
fficulty of capturing dense ground truth. In this paper, we tackle this problem
by tracking pixels through densely sampled space-time volumes recorded with a hi
gh-speed video camera. Our model exploits the linearity of small motions and rea
sons about occlusions from multiple frames. Using our technique, we are able to
establish accurate reference flow fields outside the laboratory in natural envir
onments. Besides, we show how our predictions can be used to augment the input i
mages with realistic motion blur. We demonstrate the quality of the produced flo
w fields on synthetic and real-world datasets. Finally, we collect a novel chall
enging optical flow dataset by applying our technique on data from a high-speed
camera and analyze the performance of the state-of-the-art in optical flow under
various levels of motion blur.
**********

Cross-View Image Matching for Geo-Localization in Urban Environments
Yicong Tian, Chen Chen, Mubarak Shah; Proceedings of the IEEE Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2017, pp. 3608-3616
In this paper, we address the problem of cross-view image geo-localization. Spec
ifically, we aim to estimate the GPS location of a query street view image by fi
nding the matching images in a reference database of geo-tagged bird's eye view
images, or vice versa. To this end, we present a new framework for cross-view im
age geo-localization by taking advantage of the tremendous success of deep convo
lutional neural networks (CNNs) in image classification and object detection. F
irst, we employ the Faster R-CNN to detect buildings in the query and reference
images. Next, for each building in the query image, we retrieve the k nearest ne
ighbors from the reference buildings using a Siamese network trained on both pos
itive matching image pairs and negative pairs. To find the correct NN for each q
uery building, we develop an efficient multiple nearest neighbors matching metho
d based on dominant sets. We evaluate the proposed framework on a new dataset th
at consists of pairs of street view and bird's eye view images. Experimental res
ults show that the proposed method achieves better geo-localization accuracy tha
n other approaches and is able to generalize to images at unseen locations.
**********

Improving Pairwise Ranking for Multi-Label Image Classification
Yuncheng Li, Yale Song, Jiebo Luo; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 3617-3625
Learning to rank has recently emerged as an attractive technique to train deep c onvolutional neural networks for various computer vision tasks. Pairwise ranking , in particular, has been successful in multi-label image classification, achiev ing state-of-the-art results on various benchmarks. However, most existing appro aches use the hinge loss to train their models, which is non-smooth and thus is difficult to optimize especially with deep networks. Furthermore, they employ si mple heuristics, such as top-k or thresholding, to determine which labels to inc lude in the output from a ranked list of labels, which limits their use in the r eal-world setting. In this work, we propose two techniques to improve pairwise r anking based multi-label image classification by solving the aforementioned prob lems: (1) we propose a novel loss function for pairwise ranking, which is smooth everywhere; and (2) we incorporate a label decision module into the model, esti mating the optimal confidence thresholds for each visual concept. We provide the oretical analyses of our loss function from the point of view of the Bayes consi stency and risk minimization, and show its benefit over existing pairwise rankin g formulations. We also demonstrate the effectiveness of our approach on two lar ge-scale datasets, NUS-WIDE and MS-COCO, achieving the best reported result in t he literature.
**********

Webly Supervised Semantic Segmentation
Bin Jin, Maria V. Ortiz Segovia, Sabine Susstrunk; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3626-3635
We propose a weakly supervised semantic segmentation algorithm that uses image t ags for supervision. We apply the tags in queries to collect three sets of web i mages, which encode the clean foregrounds, the common back- grounds, and realist ic scenes of the classes. We introduce a novel three-stage training pipeline to progressively learn semantic segmentation models. We first train and refine a cl ass-specific shallow neural network to obtain segmentation masks for each class. The shallow neural networks of all classes are then assembled into one deep con volutional neural network for end-to-end training and testing. Experiments show that our method notably outperforms previous state-of-the-art weakly supervised semantic segmentation approaches on the PASCAL VOC 2012 segmentation bench- mark . We further apply the class-specific shallow neural networks to object segmenta tion and obtain excellent results.
**********

Self-Supervised Video Representation Learning With Odd-One-Out Networks
Basura Fernando, Hakan Bilen, Efstratios Gavves, Stephen Gould; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3636-3645
We propose a new self-supervised CNN pre-training technique based on a novel aux iliary task called odd-one-out learning. In this task, the machine is asked to i dentify the unrelated or odd element from a set of otherwise related elements. W e apply this technique to self-supervised video representation learning where we sample subsequences from videos and ask the network to learn to predict the odd video subsequence. The odd video subsequence is sampled such that it has wrong temporal order of frames while the even ones have the correct temporal order. Th erefore, to generate a odd-one-out question no manual annotation is required. Ou r learning machine is implemented as multi-stream convolutional neural network, which is learned end-to-end. Using odd-one-out networks, we learn temporal repre sentations for videos that generalizes to other related tasks such as action rec ognition. On action classification, our method obtains 60.3% on the UCF101 data set using only UCF101 data for training which is approximately 10% better than c urrent state-of-the-art self-supervised learning methods. Similarly, on HMDB51 d ataset we outperform self-supervised state-of-the art methods by 12.7% on action classification task.
**********

Fast Video Classification via Adaptive Cascading of Deep Models

Haichen Shen, Seungyeop Han, Matthai Philipose, Arvind Krishnamurthy; Proceeding
s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
, pp. 3646-3654
Recent advances have enabled "oracle" classifiers that can classify across many
classes and input distributions with high accuracy without retraining. However,
these classifiers are relatively heavyweight, so that applying them to classify
video is costly. We show that day-to-day video exhibits highly skewed class dist
ributions over the short term, and that these distributions can be classified by
much simpler models. We formulate the problem of detecting the short-term skews
online and exploiting models based on it as a new sequential decision making pr
oblem dubbed the Online Bandit Problem, and present a new algorithm to solve it.
When applied to recognizing faces in TV shows and movies, we realize end-to-end
classification speedups of 2.4-7.8x/2.6-11.2x (on GPU/CPU) relative to a state-
of-the-art convolutional neural network, at competitive accuracy.
**********

Non-Contact Full Field Vibration Measurement Based on Phase-Shifting
Hiroyuki Kayaba, Yuji Kokumai; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2017, pp. 3655-3663
Vibration measurement systems are widely used in the industry.  A variety of vib
ration measurement techniques are proposed, including methods using an accelerat
ion sensor, a laser displacement meter, and tracking a marker using a camera. Ho
wever, these methods have limitations that allow only one point to be measured a
nd require markers. We present a novel, non-contact full field joint measurement
technique both of vibrations and shape based on phase-shifting. Our key idea is
to acquire the frequency of vibrating objects using FFT to analyze the phase-sh
ift error of vibrating objects. Our proposed algorithm estimates the phase-shift
error by iterating frame-to-frame optimization and pixel-to-pixel optimization.
A feature of our approach is to measure the surface of vibrating at different f
requencies without markers or texture in full fields.  Our developed system is a
low cost system, which is composed of a digital-light-processing (DLP) projecto
r and camera (100 frames per second). The results of our experiments show that l
ow frequency vibration of objects can be measured with high accuracy in non-cont
act.  Also, reconstruction of the vibrating object surface can be performed with
high accuracy.
**********

FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segment
ation of Generic Objects in Videos
Suyog Dutt Jain, Bo Xiong, Kristen Grauman; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3664-3673
We propose an end-to-end learning framework for segmenting generic objects in vi
deos. Our method learns to combine appearance and motion information to produce
pixel level segmentation masks for all prominent objects in videos. We formulate
this task as a structured prediction problem and design a two-stream fully conv
olutional neural network which fuses together motion and appearance in a unified
framework. Since large-scale video datasets with pixel level segmentations are
problematic, we show how to bootstrap weakly annotated videos together with exis
ting image recognition datasets for training. Through experiments on three chall
enging video segmentation benchmarks, our method substantially improves the stat
e-of-the-art for segmenting generic (unseen) objects. Code and pre-trained model
s are available on the project website.
**********

Variational Autoencoded Regression: High Dimensional Regression of Visual Data o
n Complex Manifold
YoungJoon Yoo, Sangdoo Yun, Hyung Jin Chang, Yiannis Demiris, Jin Young Choi; Pr
oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP
R), 2017, pp. 3674-3683
This paper proposes a new high dimensional regression method by merging Gaussian
process regression into a variational autoencoder framework. In contrast to oth
er regression methods, the proposed method focuses on the case where output resp
onses are on a complex high dimensional manifold, such as images. Our contributi

ons are summarized as follows: (i) A new regression method estimating high dimen sional image responses, which is not handled by existing regression algorithms, is proposed. (ii) The proposed regression method introduces a strategy to learn the latent space as well as the encoder and decoder so that the result of the re gressed response in the latent space coincide with the corresponding response in the data space. (iii) The proposed regression is embedded into a generative mod el, and the whole procedure is developed by the variational autoencoder framewor k. We demonstrate the robustness and effectiveness of our method through a numbe r of experiments on various visual data regression problems.
**********

Temporal Action Localization by Structured Maximal Sums
Zehuan Yuan, Jonathan C. Stroud, Tong Lu, Jia Deng; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3684-3692
We address the problem of temporal action localization in videos. We pose action localization as a structured prediction over arbitrary-length temporal windows, where each window is scored as the sum of frame-wise classification scores. Add itionally, our model classifies the start, middle, and end of each action as sep arate components, allowing our system to explicitly model each action's temporal evolution and take advantage of informative temporal dependencies present in th at structure. In this framework, we localize actions by searching for the struct ured maximal sum, a problem for which we develop a novel, provably-efficient alg orithmic solution. The frame-wise classification scores are computed using featu res from a deep Convolutional Neural Network (CNN), which are trained end-to-end to directly optimize for a novel structured objective. We evaluate our system o n the THUMOS '14 action detection benchmark and achieve competitive performance.
**********

Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs
Martin Simonovsky, Nikos Komodakis; Proceedings of the IEEE Conference on Comput er Vision and Pattern Recognition (CVPR), 2017, pp. 3693-3702
A number of problems can be formulated as prediction on graph-structured data. I n this work, we generalize the convolution operator from regular grids to arbitr ary graphs while avoiding the spectral domain, which allows us to handle graphs of varying size and connectivity. To move beyond a simple diffusion, filter weig hts are conditioned on the specific edge labels in the neighborhood of a vertex. Together with the proper choice of graph coarsening, we explore constructing de ep neural networks for graph classification. In particular, we demonstrate the g enerality of our formulation in point cloud classification, where we set the new state of the art, and on a graph classification dataset, where we outperform ot her deep learning approaches.
**********

Synthesizing Normalized Faces From Facial Identity Features
Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, Will iam T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Patter n Recognition (CVPR), 2017, pp. 3703-3712
We present a method for synthesizing a frontal, neutral-expression image of a pe rson's face, given an input face photograph. This is achieved by learning to gen erate facial landmarks and textures from features extracted from a facial-recogn ition network. Unlike previous generative approaches, our encoding feature vecto r is largely invariant to lighting, pose, and facial expression. Exploiting this invariance, we train our decoder network using only frontal, neutral-expression photographs. Since these photographs are well aligned, we can decompose them in to a sparse set of landmark points and aligned texture maps. The decoder then pr edicts landmarks and textures independently and combines them using a differenti able image warping operation. The resulting images can be used for a number of a pplications, such as analyzing facial attributes, exposure and white balance adj ustment, or creating a 3-D avatar.
**********

Task-Driven Dynamic Fusion: Reducing Ambiguity in Video Description
Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, Qi Tian; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

2017, pp. 3713-3721
Integrating complementary features from multiple channels is expected to solve t
he description ambiguity problem in video captioning, whereas inappropriate fusi
on strategies often harm rather than help the performance. Existing static fusio
n methods in video captioning such as concatenation and summation cannot attend
to appropriate feature channels, thus fail to adaptively support the recognition
 of various kinds of visual entities such as actions and objects. This paper con
tributes to: 1)The first in-depth study of the weakness inherent in data-driven
static fusion methods for video captioning. 2) The establishment of a task-drive
n dynamic fusion (TDDF) method. It can adaptively choose different fusion patter
ns according to model status. 3) The improvement of video captioning. Extensive
experiments conducted on two well-known benchmarks demonstrate that our dynamic
fusion method outperforms the state-of-the-art results on MSVD with METEOR score
s 0.333, and achieves superior METEOR scores 0.278 on MSR-VTT-10K. Compared to s
ingle features, the relative improvement derived from our fusion method are 10.0
% and 5.7% respectively on two datasets.
**********

Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks
Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, Dilip Kris
hnan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2017, pp. 3722-3731
Collecting well-annotated image datasets to train modern machine learning algori
thms is prohibitively expensive for many tasks. One appealing alternative is ren
dering synthetic data where ground-truth annotations are generated automatically
. Unfortunately, models trained purely on rendered images fail to generalize to
real images. To address this shortcoming, prior work introduced unsupervised dom
ain adaptation algorithms that have tried to either map representations between
the two domains, or learn to extract features that are domain-invariant. In this
 work, we approach the problem in a new light by learning  in an unsupervised ma
nner a transformation in the pixel space from one domain to the other. Our gener
ative adversarial network (GAN)-based method adapts source-domain images to appe
ar as if drawn from the target domain. Our approach not only produces plausible
samples, but also outperforms the state-of-the-art on a number of unsupervised d
omain adaptation scenarios by large margins. Finally, we demonstrate that the ad
aptation process generalizes to object classes unseen during training.
**********

Simultaneous Visual Data Completion and Denoising Based on Tensor Rank and Total
 Variation Minimization and Its Primal-Dual Splitting Algorithm
Tatsuya Yokota, Hidekata Hontani; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2017, pp. 3732-3740
Tensor completion has attracted attention because of its promising ability and g
enerality.  However, there are few studies on noisy scenarios which directly sol
ve an optimization problem consisting of a "noise inequality constraint".  In th
is paper, we propose a new tensor completion and denoising model including tenso
r total variation and tensor nuclear norm minimization with a range of values an
d noise inequalities.  Furthermore, we developed its solution algorithm based on
 a primal-dual splitting method, which is computationally efficient as compared
to tensor decomposition based non-convex optimization.  Lastly, extensive experi
ments demonstrated the advantages of the proposed method for visual data retriev
al such as for color images, movies, and 3D-volumetric data.
**********

Point to Set Similarity Based Deep Feature Learning for Person Re-Identification
Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, Nanning Zheng; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,
pp. 3741-3750
Person re-identification (Re-ID) remains a challenging problem due to significan
t appearance changes caused by variations in  view angle, background clutter, il
lumination condition and mutual occlusion. To address these issues, conventional
 methods usually focus on proposing robust feature representation or learning me
tric transformation based on pairwise similarity, using Fisher-type criterion. T

he recent development in deep learning based approaches address the two processes in a joint fashion and have achieved promising progress. One of the key issues for deep learning based person Re-ID is the selection of proper similarity comparison criteria, and the performance of learned features using existing criterion based on pairwise similarity is still limited, because only P2P distances are mostly considered. In this paper, we present a novel person Re-ID method based on P2S similarity comparison. The P2S metric can jointly minimize the intra-class distance and maximize the inter-class distance, while back-propagating the gradient to optimize parameters of the deep model. By utilizing our proposed P2S metric, the learned deep model can effectively distinguish different persons by learning discriminative and stable feature representations. Comprehensive experimental evaluations on 3DPeS, CUHK01, PRID2011 and Market1501 datasets demonstrate the advantages of our method over the state-of-the-art approaches.
**********

Gated Feedback Refinement Network for Dense Image Labeling
Md Amirul Islam, Mrigank Rochan, Neil D. B. Bruce, Yang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3751-3759
Effective integration of local and global contextual information is crucial for dense labeling problems. Most existing methods based on an encoder-decoder architecture simply concatenate features from earlier layers to obtain higher-frequency details in the refinement stages. However, there are limits to the quality of refinement possible if ambiguous information is passed forward. In this paper we propose Gated Feedback Refinement Network (G-FRNet), an end-to-end deep learning framework for dense labeling tasks that addresses this limitation of existing methods. Initially, G-FRNet makes a coarse prediction and then it progressively refines the details by efficiently integrating local and global contextual information during the refinement stages. We introduce gate units that control the information passed forward in order to filter out ambiguity. Experiments on three challenging dense labeling datasets (CamVid, PASCAL VOC 2012, and Horse-Cow Parsing) show the effectiveness of our method. Our proposed approach achieves state-of-the-art results on the CamVid and Horse-Cow Parsing datasets, and produces competitive results on the PASCAL VOC 2012 dataset.
**********

Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders
Xin Yu, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3760-3768
Most of the conventional face hallucination methods assume the input image is sufficiently large and aligned, and all require the input image to be noise-free. Their performance degrades drastically if the input image is tiny, unaligned, and contaminated by noise. In this paper, we introduce a novel transformative discriminative autoencoder to 8X super-resolve unaligned noisy and tiny (16X16) low-resolution face images. In contrast to encoder-decoder based autoencoders, our method uses decoder-encoder-decoder networks. We first employ a transformative discriminative decoder network to upsample and denoise simultaneously. Then we use a transformative encoder network to project the intermediate HR faces to aligned and noise-free LR faces. Finally, we use the second decoder to generate hallucinated HR images. Our extensive evaluations on a very large face dataset show that our method achieves superior hallucination results and outperforms the state-of-the-art by a large margin of 1.82dB PSNR.
**********

Learning Dynamic Guidance for Depth Image Enhancement
Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3769-3778
The depth images acquired by consumer depth sensors (e.g., Kinect and ToF) usually are of low resolution and insufficient quality. One natural solution is to incorporate with high resolution RGB camera for exploiting their statistical correlation. However, most existing methods are intuitive and limited in characterizi

ng the complex and dynamic dependency between intensity and depth images. To add ress these limitations, we propose a weighted analysis representation model for guided depth image enhancement, which advances the conventional methods in two a spects: (i) task driven learning and (ii) dynamic guidance. First, we generalize the analysis representation model by including a guided weight function for dep endency modeling. And the task-driven learning formulation is introduced to obta in the optimized guidance tailored to specific enhancement task. Second, the dep th image is gradually enhanced along with the iterations, and thus the guidance should also be dynamically adjusted to account for the updating of depth image. To this end, stage-wise parameters are learned for dynamic guidance. Experiment s on guided depth image upsampling and noisy depth image restoration validate th e effectiveness of our method.

**********

3D Shape Segmentation With Projective Convolutional Networks

Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, Siddhartha Chaudhuri; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 3779-3788

This paper introduces a deep architecture for segmenting 3D objects into their l abeled semantic parts. Our architecture combines image-based Fully Convolutiona l Networks (FCNs) and surface-based Conditional Random Fields (CRFs) to yield co herent segmentations of 3D shapes. The image-based FCNs are used for efficient v iew-based reasoning about 3D object parts. Through a special projection layer, F CN outputs are effectively aggregated across multiple views and scales, then are projected onto the 3D object surfaces. Finally, a surface-based CRF combines th e projected outputs with geometric consistency cues to yield coherent segmentati ons. The whole architecture (multi-view FCNs and CRF) is trained end-to-end. Our approach significantly outperforms the existing state-of-the-art methods in the currently largest segmentation benchmark (ShapeNet). Finally, we demonstrate pr omising segmentation results on noisy 3D shapes acquired from consumer-grade dep th cameras.

**********

Deep Image Harmonization

Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, Ming-Hsuan Yang ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3789-3797

Compositing is one of the most common operations in photo editing. To generate r ealistic composites, the appearances of foreground and background need to be adj usted to make them compatible. Previous approaches to harmonize composites have focused on learning statistical relationships between hand-crafted appearance fe atures of the foreground and background, which is unreliable especially when the contents in the two layers are vastly different. In this work, we propose an en d-to-end deep convolutional neural network for image harmonization, which can ca pture both the context and semantic information of the composite images during h armonization. We also introduce an efficient way to collect large-scale and high -quality training data that can facilitate the training process. Experiments on the synthesized dataset and real composite images show that the proposed network outperforms previous state-of-the-art methods.

**********

Matrix Tri-Factorization With Manifold Regularizations for Zero-Shot Learning

Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen, Jingkuan Song; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 3798-3807

Zero-shot learning (ZSL) aims to recognize objects of unseen classes with availa ble training data from another set of seen classes. Existing solutions are focus ed on exploring knowledge transfer via an intermediate semantic embedding (e.g.s , attributes) shared between seen and unseen classes. In this paper, we propose a novel projection framework based on matrix tri-factorization with manifold reg ularizations. Specifically, we learn the semantic embedding projection by decomp osing the visual feature matrix under the guidance of semantic embedding and cla ss label matrices. By additionally introducing manifold regularizations on visua

l data and semantic embeddings, the learned projection can effectively captures the geometrical manifold structure residing in both visual and semantic spaces. To avoid the projection domain shift problem, we devise an effective prediction scheme by exploiting the test-time manifold structure. Extensive experiments on four benchmark datasets show that our approach significantly outperforms the sta te-of-the-arts, yielding an average improvement  ratio by 7.4% and 31.9% for the recognition and retrieval task, respectively.
**********

Spatio-Temporal Alignment of Non-Overlapping Sequences From Independently Pannin g Cameras
Seyed Morteza Safdarnejad, Xiaoming Liu; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2017, pp. 3808-3816
This paper addresses the problem of spatio-temporal alignment of multiple video sequences. We identify and tackle a novel scenario of this problem referred to a s Nonoverlapping Sequences (NOS). NOS are captured by multiple freely panning ha ndheld cameras whose field of views (FOV) might have no direct spatial overlap. With the popularity of mobile sensors, NOS rise when multiple cooperative users capture a public event to create a panoramic video, or when consolidating multip le footages of an incident into a single video. To tackle this novel scenario, w e first spatially align the sequences by reconstructing the background of each s equence and registering these backgrounds, even if the backgrounds are not overl apping. Given the spatial alignment, we temporally synchronize the sequences, su ch that the trajectories of moving objects (e.g., cars or pedestrians) are consi stent across sequences. Experimental results demonstrate the performance of our algorithm in this novel and challenging scenario, quantitatively and qualitative ly.
**********

Learning Fully Convolutional Networks for Iterative Non-Blind Deconvolution
Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson W. H. Lau, Ming-Hsuan Yang; Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ), 2017, pp. 3817-3825
In this paper, we propose a fully convolutional network for iterative non-blind deconvolution. We decompose the non-blind deconvolution problem into image denoi sing and image deconvolution. We train a FCNN to remove noise in the gradient do main and use the learned gradients to guide the image deconvolution step. In con trast to the existing deep neural network based methods, we iteratively deconvol ve the blurred images in a multi-stage framework. The proposed method is able to learn an adaptive image prior, which keeps both local (details) and global (str uctures) information. Both quantitative and qualitative evaluations on the bench mark datasets demonstrate that the proposed method performs favorably against st ate-of-the-art algorithms in terms of quality and speed.
**********

Seeing Into Darkness: Scotopic Visual Recognition
Bo Chen, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2017, pp. 3826-3835
Images are formed by counting how many photons traveling from a given set of dir ections hit an image sensor during a given time interval. When photons are few a nd far in between, the concept of `image' breaks down and it is best to consider directly the flow of photons. Computer vision in this regime, which we call `sc otopic', is radically different from the classical image-based paradigm in that visual computations (classification, control, search) have to take place while t he stream of photons is captured and decisions may be taken as soon as enough in formation is available. The scotopic regime is important for biomedical imaging, security, astronomy and many other fields. Here we develop a framework that all ows a machine to classify objects with as few photons as possible, while maintai ning the error rate below an acceptable threshold. A dynamic and asymptotically optimal speed-accuracy tradeoff is a key feature of this framework. We propose a nd study an algorithm to optimize the tradeoff of a convolutional network direct ly from lowlight images and evaluate on simulated images from standard datasets. Surprisingly, scotopic systems can achieve comparable classification performanc

e as traditional vision systems while using less than 0.1% of the photons in a c onventional image. In addition, we demonstrate that our algorithms work even whe n the illuminance of the environment is unknown and varying. Last, we outline a spiking neural network coupled with photon-counting sensors as a power-efficient hardware realization of scotopic algorithms.
**********

Distinguishing the Indistinguishable: Exploring Structural Ambiguities via Geode sic Context
Qingan Yan, Long Yang, Ling Zhang, Chunxia Xiao; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3836-3844
A perennial problem in structure from motion (SfM) is visual ambiguity posed by repetitive structures. Recent disambiguating algorithms infer ambiguities mainly via explicit background context, thus face limitations in highly ambiguous scen es which are visually indistinguishable. Instead of analyzing local visual infor mation, we propose a novel algorithm for SfM disambiguation that explores the gl obal topology as encoded in photo collections. An important adaptation of this w ork is to approximate the available imagery using a manifold of viewpoints. We n ote that, while ambiguous images appear deceptively similar in appearance, they are actually located far apart on geodesics. We establish the manifold by adapti vely identifying cameras with adjacent viewpoint, and detect ambiguities via a n ew measure, geodesic consistency. We demonstrate the accuracy and efficiency of the proposed approach on a range of complex ambiguity datasets, even including t he challenging scenes without background conflicts.
**********

Learning an Invariant Hilbert Space for Domain Adaptation
Samitha Herath, Mehrtash Harandi, Fatih Porikli; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3845-3854
This paper introduces a learning scheme to construct a Hilbert space (i.e., a ve ctor space along its inner product) to address both unsupervised and semi-superv ised domain adaptation problems. This is achieved by learning projections from e ach domain to a latent space along the Mahalanobis metric of the latent space to simultaneously minimizing a notion of domain variance while maximizing a measur e of discriminatory power. In particular, we make use of the Riemannian optimiza tion techniques to match statistical properties (e.g., first and second order st atistics) between samples projected into the latent space from different domains . Upon availability of class labels, we further deem samples sharing the same la bel to form more compact clusters while pulling away samples coming from differe nt classes.We extensively evaluate and contrast our proposal against state-of-th e-art methods for the task of visual domain adaptation using both handcrafted an d deep-net features. Our experiments show that even with a simple nearest neighb or classifier, the proposed method can outperform several state-of-the-art metho ds benefitting from more involved classification schemes.
**********

Removing Rain From Single Images via a Deep Detail Network
Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, John Paisley; Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2017, pp. 3855-3863
We propose a new deep network architecture for removing rain streaks from indivi dual images based on the deep convolutional neural network (CNN). Inspired by th e deep residual network (ResNet) that simplifies the learning process by changin g the mapping form, we propose a deep detail network to directly reduce the mapp ing range from input to output, which makes the learning process easier. To furt her improve the de-rained result, we use a priori image domain knowledge by focu sing on high frequency detail during training, which removes background interfer ence and focuses the model on the structure of rain in images. This demonstrates that a deep architecture not only has benefits for high-level vision tasks but also can be used to solve low-level imaging problems. Though we train the networ k on synthetic data, we find that the learned network generalizes well to real-w orld test images. Experiments show that the proposed method significantly outper forms state-of-the-art methods on both synthetic and real-world images in terms

of both qualitative and quantitative measures. We discuss applications of this s
tructure to denoising and JPEG artifact reduction at the end of the paper.
**********

Binarized Mode Seeking for Scalable Visual Pattern Discovery
Wei Zhang, Xiaochun Cao, Rui Wang, Yuanfang Guo, Zhineng Chen; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3
864-3872
This paper studies visual pattern discovery in large-scale image collections via
 binarized mode seeking, where images can only be represented as binary codes fo
r efficient storage and computation. We address this problem from the perspectiv
e of binary space mode seeking. First, a binary mean shift (bMS) is proposed to
discover frequent patterns via mode seeking directly in binary space. The binomi
al-based kernel and binary constraint are introduced for binarized analysis. Sec
ond, we further extend bMS to a more general form, namely contrastive binary mea
n shift (cbMS), which maximizes the contrastive density in binary space, for fin
ding informative patterns that are both frequent and discriminative for the data
set. With the binarized algorithm and optimization, our methods demonstrate sign
ificant computation (50X) and storage (32X) improvement compared to standard tec
hniques operating in Euclidean space, while the performance does not largely deg
enerate. Furthermore, cbMS discovers more informative patterns by suppressing lo
w discriminative modes. We evaluate our methods on both annotated ILSVRC (1M ima
ges) and un-annotated blind Flickr (10M images) datasets with million scale imag
es, which demonstrates both the scalability and effectiveness of our algorithms
for discovering frequent and informative patterns in large scale collection.
**********

Fast Person Re-Identification via Cross-Camera Semantic Binary Transformation
Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, Ling Shao; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3873-3882
Numerous methods have been proposed for person re-identification, most of which
however neglect the matching efficiency. Recently, several hashing based approac
hes have been developed to make re-identification more scalable for large-scale
gallery sets. Despite their efficiency, these works ignore cross-camera variatio
ns, which severely deteriorate the final matching accuracy. To address the above
 issues, we propose a novel hashing based method for fast person re-identificati
on, namely Cross-camera Semantic Binary Transformation (CSBT). CSBT aims to tran
sform original high-dimensional feature vectors into compact identity-preserving
 binary codes. To this end, CSBT first employs a subspace projection to mitigate
 cross-camera variations, by maximizing intra-person similarities and inter-pers
on discrepancies. Subsequently, a binary coding scheme is proposed via seamlessl
y incorporating both the semantic pairwise relationships and local affinity info
rmation. Finally, a joint learning framework is proposed for simultaneous subspa
ce projection learning and binary coding based on discrete alternating optimizat
ion. Experimental results on four benchmarks clearly demonstrate the superiority
 of CSBT over the state-of-the-art methods.
**********

Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring
Seungjun Nah, Tae Hyun Kim, Kyoung Mu Lee; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3883-3891
Non-uniform blind deblurring for general dynamic scenes is a challenging compute
r vision problem as blurs arise not only from multiple object motions but also f
rom camera shake, scene depth variation. To remove these complicated motion blur
s, conventional energy optimization based methods rely on simple assumptions suc
h that blur kernel is partially uniform or locally linear. Moreover, recent mach
ine learning based methods also depend on synthetic blur datasets generated unde
r these assumptions. This makes conventional deblurring methods fail to remove b
lurs where blur kernel is difficult to approximate or parameterize (e.g. object
motion boundaries). In this work, we propose a multi-scale convolutional neural
network that restores sharp images in an end-to-end manner where blur is caused
by various sources. Together, we present multi-scale loss function that mimics c
onventional coarse-to-fine approaches. Furthermore, we propose a new large-scale

dataset that provides pairs of realistic blurry image and the corresponding gro
und truth sharp image that are obtained by a high-speed camera. With the propose
d model trained on this dataset, we demonstrate empirically that our method achi
eves the state-of-the-art performance in dynamic scene deblurring not only quali
tatively, but also quantitatively.
**********

Deep Crisp Boundaries

Yupei Wang, Xin Zhao, Kaiqi Huang; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2017, pp. 3892-3900

Edge detection had made significant progress with the help of deep Convolutional
Networks (ConvNet). ConvNet based edge detectors approached human level perform
ance on standard benchmarks. We provide a systematical study of these detector o
utputs, and show that they failed to accurately localize edges, which can be adv
ersarial for tasks that require crisp edge inputs. In addition, we propose a nov
el refinement architecture to address the challenging problem of learning a cris
p edge detector using ConvNet. Our method leverages a top-down backward refineme
nt pathway, and progressively increases the resolution of feature maps to genera
te crisp edges. Our results achieve promising performance on BSDS500, surpassing
human accuracy when using standard criteria, and largely outperforming state-of
-the-art methods when using more strict criteria. We further demonstrate the ben
efit of crisp edge maps for estimating optical flow and generating object propos
als.
**********

Learning Multifunctional Binary Codes for Both Category and Attribute Oriented R
etrieval Tasks

Haomiao Liu, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3901-3910

In this paper we propose a unified framework to address multiple realistic image
retrieval tasks concerning both category and attributes. Considering the scale
of modern datasets, hashing is favorable for its low complexity. However, most e
xisting hashing methods are designed to preserve one single kind of similarity,
thus incapable of dealing with the different tasks simultaneously. To overcome t
his limitation, we propose a new hashing method, named Dual Purpose Hashing (DPH
), which jointly preserves the category and attribute similarities by exploiting
the convolutional networks (CNN) to hierarchically capture the correlations bet
ween category and attributes. Since images with both category and attribute labe
ls are scarce, our method is designed to take the abundant partially labelled im
ages on the Internet as training inputs. With such a framework, the binary codes
of new-coming images can be readily obtained by quantizing the network outputs
of a binary-like layer, and the attributes can be recovered from the codes easil
y. Experiments on two large-scale datasets show that our dual purpose hash codes
can achieve comparable or even better performance than those state-of-the-art m
ethods specifically designed for each individual retrieval task, while being mor
e compact than the compared methods.
**********

Generative Face Completion

Yijun Li, Sifei Liu, Jimei Yang, Ming-Hsuan Yang; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3911-3919

In this paper, we propose an effective face completion algorithm using a deep ge
nerative model. Different from well-studied background completion, the face comp
letion task is more challenging as it often requires to generate semantically ne
w pixels for the missing key components (e.g., eyes and mouths) that contain lar
ge appearance variations. Unlike existing nonparametric algorithms that search f
or patches to synthesize, our algorithm directly generates contents for missing
regions based on a neural network. The model is trained with a combination of a
reconstruction loss, two adversarial losses and a semantic parsing loss, which e
nsures pixel faithfulness and local-global contents consistency. With extensive
experimental results, we demonstrate qualitatively and quantitatively that our m
odel is able to deal with a large area of missing pixels in arbitrary shapes and
generate realistic face completion results.

**********
Diversified Texture Synthesis With Feed-Forward Networks
Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3920-3928
Recent progresses on deep discriminative and generative modeling have shown promising results on texture synthesis. However, existing feed-forward based methods trade off generality for efficiency, which suffer from many issues, such as shortage of generality (i.e., build one network per texture), lack of diversity (i.e., always produce visually identical output) and suboptimality (i.e., generate less satisfying visual effects). In this work, we focus on solving these issues for improved texture synthesis. We propose a deep generative feed-forward network which enables efficient synthesis of multiple textures within one single network and meaningful interpolation between them. Meanwhile, a suite of important techniques are introduced to achieve better convergence and diversity. With extensive experiments, we demonstrate the effectiveness of the proposed model and techniques for synthesizing a large number of textures and show its applications with the stylization.
**********
Learning Deep CNN Denoiser Prior for Image Restoration
Kai Zhang, Wangmeng Zuo, Shuhang Gu, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3929-3938
Model-based optimization methods and discriminative learning methods have been the two dominant strategies for solving various inverse problems in low-level vision. Typically, those two kinds of methods have their respective merits and drawbacks, e.g., model-based optimization methods are flexible for handling different inverse problems but are usually time-consuming with sophisticated priors for the purpose of good performance; in the meanwhile, discriminative learning methods have fast testing speed but their application range is greatly restricted by the specialized task. Recent works have revealed that, with the aid of variable splitting techniques, denoiser prior can be plugged in as a modular part of model-based optimization methods to solve other inverse problems (e.g., deblurring). Such an integration induces considerable advantage when the denoiser is obtained via discriminative learning. However, the study of integration with fast discriminative denoiser prior is still lacking. To this end, this paper aims to train a set of fast and effective CNN (convolutional neural network) denoisers and integrate them into model-based optimization method to solve other inverse problems. Experimental results demonstrate that the learned set of denoisers can not only achieve promising Gaussian denoising results but also can be used as prior to deliver good performance for various low-level vision applications.
**********
Fast Multi-Frame Stereo Scene Flow With Motion Segmentation
Tatsunori Taniai, Sudipta N. Sinha, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3939-3948
We propose a new multi-frame method for efficiently computing scene flow (dense depth and optical flow) and camera ego-motion for a dynamic scene observed from a moving stereo camera rig. Our technique also segments out moving objects from the rigid scene. In our method, we first estimate the disparity map and the 6-DOF camera motion using stereo matching and visual odometry. We then identify regions inconsistent with the estimated camera motion and compute per-pixel optical flow only at these regions. This flow proposal is fused with the camera motion-based flow proposal using fusion moves to obtain the final optical flow and motion segmentation. This unified framework benefits all four tasks -- stereo, optical flow, visual odometry and motion segmentation leading to overall higher accuracy and efficiency. Our method is currently ranked third on the KITTI 2015 scene flow benchmark. Furthermore, our CPU implementation runs in 2-3 seconds per frame which is 1-3 orders of magnitude faster than the top six methods. We also report a thorough evaluation on challenging Sintel sequences with fast camera and object motion, where our method consistently outperforms OSF [Menze2015], which is currently ranked second on the KITTI benchmark.

**********
DeepPermNet: Visual Permutation Learning
Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3949-3957
We present a principled approach to uncover the structure of visual data by solving a novel deep learning task coined visual permutation learning. The goal of this task is to find the permutation that recovers the structure of data from shuffled versions of it. In the case of natural images, this task boils down to recovering the original image from patches shuffled by an unknown permutation matrix. Unfortunately, permutation matrices are discrete, thereby posing difficulties for gradient-based methods. To this end, we resort to a continuous approximation of these matrices using doubly-stochastic matrices which we generate from standard CNN predictions using Sinkhorn iterations. Unrolling these iterations in a Sinkhorn network layer, we propose DeepPermNet, an end-to-end CNN model for this task. The utility of DeepPermNet is demonstrated on two challenging computer vision problems, namely, (i) relative attributes learning and (ii) self-supervised representation learning. Our results show state-of-the-art performance on the Public Figures and OSR benchmarks for (i) and on the classification and segmentation tasks on the PASCAL VOC dataset for (ii).
**********
Light Field Blind Motion Deblurring
Pratul P. Srinivasan, Ren Ng, Ravi Ramamoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3958-3966
We study the problem of deblurring light fields of general 3D scenes captured under 3D camera motion and present both theoretical and practical contributions. By analyzing the motion-blurred light field in the primal and Fourier domains, we develop intuition into the effects of camera motion on the light field, show the advantages of capturing a 4D light field instead of a conventional 2D image for motion deblurring, and derive simple analytical methods of motion deblurring in certain cases. We then present an algorithm to blindly deblur light fields of general scenes without any estimation of scene geometry, and demonstrate that we can recover both the sharp light field and the 3D camera motion path of real and synthetically-blurred light fields.
**********
Wetness and Color From a Single Multispectral Image
Mihoko Shimano, Hiroki Okawa, Yuta Asano, Ryoma Bise, Ko Nishino, Imari Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3967-3975
Visual recognition of wet surfaces and their degrees of wetness is important for many computer vision applications. It can inform slippery spots on a road to autonomous vehicles, muddy areas of a trail to humanoid robots, and the freshness of groceries to us. In the past, monochromatic appearance change, the fact that surfaces darken when wet, has been modeled to recognize wet surfaces. In this paper, we show that color change, particularly in its spectral behavior, carries rich information about a wet surface. We derive an analytical spectral appearance model of wet surfaces that expresses the characteristic spectral sharpening due to multiple scattering and absorption in the surface. We derive a novel method for estimating key parameters of this spectral appearance model, which enables the recovery of the original surface color and the degree of wetness from a single observation. Applied to a multispectral image, the method estimates the spatial map of wetness together with the dry spectral distribution of the surface. To our knowledge, this work is the first to model and leverage the spectral characteristics of wet surfaces to revert its appearance. We conduct comprehensive experimental validation with a number of wet real surfaces. The results demonstrate the accuracy of our model and the effectiveness of our method for surface wetness and color estimation.
**********
Seeing Invisible Poses: Estimating 3D Body Pose From Egocentric Video
Hao Jiang, Kristen Grauman; Proceedings of the IEEE Conference on Computer Visio

n and Pattern Recognition (CVPR), 2017, pp. 3976-3984
Understanding the camera wearer's activity is central to egocentric vision, yet one key facet of that activity is inherently invisible to the camera--the wearer's body pose. Prior work focuses on estimating the pose of hands and arms when they come into view, but this 1) gives an incomplete view of the full body posture, and 2) prevents any pose estimate at all in many frames, since the hands are only visible in a fraction of daily life activities. We propose to infer the "invisible pose" of a person behind the egocentric camera. Given a single video, our efficient learning-based approach returns the full body 3D joint positions for each frame. Our method exploits cues from the dynamic motion signatures of the surrounding scene--which change predictably as a function of body pose--as well as static scene structures that reveal the viewpoint (e.g., sitting vs. standing). We further introduce a novel energy minimization scheme to infer the pose sequence. It uses soft predictions of the poses per time instant together with a non-parametric model of human pose dynamics over longer windows. Our method outperforms an array of possible alternatives, including typical deep learning approaches for direct pose regression from images.
**********

Controlling Perceptual Factors in Neural Style Transfer
Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, Eli Shechtman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3985-3993
Neural Style Transfer has shown very exciting results enabling new forms of image manipulation. Here we extend the existing method to introduce control over spatial location, colour information and across spatial scale. We demonstrate how this enhances the method by allowing high-resolution controlled stylisation and helps to alleviate common failure cases such as applying ground textures to sky regions. Furthermore, by decomposing style into these perceptual factors we enable the combination of style information from multiple sources to generate new, perceptually appealing styles from existing ones. We also describe how these methods can be used to more efficiently produce large size, high-quality stylisation. Finally we show how the introduced control measures can be applied in recent methods for Fast Neural Style Transfer.
**********

Learning to Rank Retargeted Images
Yang Chen, Yong-Jin Liu, Yu-Kun Lai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3994-4002
Image retargeting techniques that adjust images into different sizes have attracted much attention recently. Objective quality assessment (OQA) of image retargeting results is often desired to automatically select the best results. Existing OQA methods output an absolute score for each retargeted image and use these scores to compare different results. Observing that it is challenging even for human subjects to give consistent scores for retargeting results of different source images, in this paper we propose a learning-based OQA method that predicts the ranking of a set of retargeted images with the same source image. We show that this more manageable task helps achieve more consistent prediction to human preference and is sufficient for most application scenarios. To compute the ranking, we propose a simple yet efficient machine learning framework that uses a General Regression Neural Network (GRNN) to model a combination of seven elaborate OQA metrics. We then propose a simple scheme to transform the relative scores output from GRNN into a global ranking. We train our GRNN model using human preference data collected in the elaborate RetargetMe benchmark and evaluate our method based on the subjective study in RetargetMe. Moreover, we introduce a further subjective benchmark to evaluate the generalizability of different OQA methods. Experimental results demonstrate that our method outperforms eight representative OQA methods in ranking prediction and has better generalizability to different datasets.
**********

Image Deblurring via Extreme Channels Prior
Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, Xiaochun Cao; Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4003-4011
Camera motion introduces motion blur, affecting many computer vision tasks. Dark Channel Prior (DCP) helps the blind deblurring on scenes including natural, face, text, and low-illumination images. However, it has limitations and is less likely to support the kernel estimation while bright pixels dominate the input image. We observe that the bright pixels in the clear images are not likely to be bright after the blur process. Based on this observation, we first illustrate this phenomenon mathematically and define it as the Bright Channel Prior (BCP). Then, we propose a technique for deblurring such images which elevates the performance of existing motion deblurring algorithms. The proposed method takes advantage of both Bright and Dark Channel Prior. This joint prior is named as extreme channels prior and is crucial for achieving efficient restorations by leveraging both the bright and dark information. Extensive experimental results demonstrate that the proposed method is more robust and performs favorably against the state-of-the-art image deblurring methods on both synthesized and natural images.
**********

Fixed-Point Factorized Networks
Peisong Wang, Jian Cheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4012-4020
In recent years, Deep Neural Networks (DNN) based methods have achieved remarkable performance in a wide range of tasks and have been among the most powerful and widely used techniques in computer vision. However, DNN-based methods are both computational-intensive and resource-consuming, which hinders the application of these methods on embedded systems like smart phones. To alleviate this problem, we introduce a novel Fixed-point Factorized Networks (FFN) for pretrained models to reduce the computational complexity as well as the storage requirement of networks. The resulting networks have only weights of -1, 0 and 1, which significantly eliminates the most resource-consuming multiply-accumulate operations (MACs). Extensive experiments on large-scale ImageNet classification task show the proposed FFN only requires one-thousandth of multiply operations with comparable accuracy.
**********

Large Margin Object Tracking With Circulant Feature Maps
Mengmeng Wang, Yong Liu, Zeyi Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4021-4029
Structured output support vector machine (SVM) based tracking algorithms have shown favorable performance recently. Nonetheless, the time-consuming candidate sampling and complex optimization limit their real-time applications. In this paper, we propose a novel large margin object tracking method which absorbs the strong discriminative ability from structured output SVM and speeds up by the correlation filter algorithm significantly. Secondly, a multimodal target detection technique is proposed to improve the target localization precision and prevent model drift introduced by similar objects or background noise. Thirdly, we exploit the feedback from high-confidence tracking results to avoid the model corruption problem. We implement two versions of the proposed tracker with the representations from both conventional hand-crafted and deep convolution neural networks (CNNs) based features to validate the strong compatibility of the algorithm. The experimental results demonstrate that the proposed tracker performs superiorly against several state-of-the-art algorithms on the challenging benchmark sequences while runs at speed in excess of 80 frames per second.
**********

Learning Residual Images for Face Attribute Manipulation
Wei Shen, Rujie Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4030-4038
Face attributes are interesting due to their detailed description of human faces. Unlike prior researches working on attribute prediction, we address an inverse and more challenging problem called face attribute manipulation which aims at modifying a face image according to a given attribute value. Instead of manipulating the whole image, we propose to learn the corresponding residual image define

d as the difference between images before and after the manipulation. In this wa y, the manipulation can be operated efficiently with modest pixel modification. The framework of our approach is based on the Generative Adversarial Network. It consists of two image transformation networks and a discriminative network. The transformation networks are responsible for the attribute manipulation and its dual operation and the discriminative network is used to distinguish the generat ed images from real images. We also apply dual learning to allow transformation networks to learn from each other. Experiments show that residual images can be effectively learned and used for attribute manipulations. The generated images r emain most of the details in attribute-irrelevant areas.

**********

One-Shot Hyperspectral Imaging Using Faced Reflectors

Tsuyoshi Takatani, Takahito Aoto, Yasuhiro Mukaigawa; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4039-4047

Hyperspectral imaging is a useful technique for various computer vision tasks su ch as material recognition. However, such technique usually requires an expensiv e and professional setup and is time-consuming because a conventional hyperspect ral image consists of a large number of observations. In this paper, we propose a novel technique of one-shot hyperspectral imaging using faced reflectors on wh ich color filters are attached. The key idea is based on the principle that each of multiple reflections on the filters has a different spectrum, which allows u s to observe multiple intensities through different spectra. Our technique can b e implemented either by a coupled mirror or a kaleidoscope geometry. Experimenta l results show that our technique is capable of accurately capturing a hyperspec tral image by using a coupled mirror setup which is readily available.

**********

Video2Shop: Exact Matching Clothes in Videos to Online Shopping Images

Zhi-Qi Cheng, Xiao Wu, Yang Liu, Xian-Sheng Hua; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4048-4056

In recent years, both online retail and video hosting service have been exponent ially grown. In this paper, a novel deep neural network, called AsymNet, is prop osed to explore a new cross-domain task, Video2Shop, targeting for matching clot hes appeared in videos to the exactly same items in online shops. For the image side, well-established methods are used to detect and extract features for cloth ing patches with arbitrary sizes. For the video side, deep visual features are e xtracted from detected object regions in each frame, and further fed into a Long Short-Term Memory (LSTM) framework for sequence modeling, which captures the te mporal dynamics in videos. To conduct exact matching between videos and online s hopping images, LSTM hidden states for videos and image features extracted from static images are jointly modeled, under the similarity network with reconfigura ble deep tree structure. Moreover, an approximate training method is proposed to achieve the efficiency when training. Extensive experiments conducted on a larg e cross-domain dataset have demonstrated the effectiveness and efficiency of the proposed AsymNet, which outperforms the state-of-the-art methods.

**********

A Novel Tensor-Based Video Rain Streaks Removal Approach via Utilizing Discrimin atively Intrinsic Priors

Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, Liang-Jian Deng, Yao Wang; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 17, pp. 4057-4066

Rain streaks removal is an important issue of the outdoor vision system and has been recently investigated extensively. In this paper, we propose a novel tensor based video rain streaks removal approach by fully considering the discriminati vely intrinsic characteristics of rain streaks and clean videos, which needs nei ther rain detection nor time-consuming dictionary learning stage. In specific, o n the one hand, rain streaks are sparse and smooth along the raindrops' directio n, and on the other hand, the clean videos possess smoothness along the rain-per pendicular direction and global and local correlation along time direction. We u se the l_1 norm to enhance the sparsity of the underlying rain, two unidirection al Total Variation (TV) regularizers to guarantee the different discriminative s

moothness, and a tensor nuclear norm and a time directional difference operator to characterize the exclusive correlation of the clean video along time. Alternation direction method of multipliers (ADMM) is employed to solve the proposed concise tensor based convex model. Experiments implemented on synthetic and real data substantiate the effectiveness and efficiency of the proposed method. Under comprehensive quantitative performance measures, our approach outperforms other state-of-the-art methods.
**********

DeshadowNet: A Multi-Context Embedding Deep Network for Shadow Removal
Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, Rynson W. H. Lau; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4067-4075
Shadow removal is a challenging task as it requires the detection/annotation of shadows as well as semantic understanding of the scene. In this paper, we propose an automatic and end-to-end deep neural network (DeshadowNet) to tackle these problems in a unified manner. DeshadowNet is designed with a multi-context architecture, where the output shadow matte is predicted by embedding information from three different perspectives. The first global network extracts shadow features from a global view. Two levels of features are derived from the global network and transferred to two parallel networks. While one extracts the appearance of the input image, the other one involves semantic understanding for final prediction. These two complementary networks generate multi-context features to obtain the shadow matte with fine local details. To evaluate the performance of the proposed method, we construct the first large scale benchmark with 3088 image pairs. Extensive experiments on two publicly available benchmarks and our large-scale benchmark show that the proposed method performs favorably against several state-of-the-art methods.
**********

Generalized Semantic Preserving Hashing for N-Label Cross-Modal Retrieval
Devraj Mandal, Kunal N. Chaudhury, Soma Biswas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4076-4084
Due to availability of large amounts of multimedia data, cross-modal matching is gaining increasing importance. Hashing based techniques provide an attractive solution to this problem when the data size is large. Different scenarios of cross-modal matching are possible, for example, data from the different modalities can be associated with a single label or multiple labels, and in addition may or may not have one-to-one correspondence. Most of the existing approaches have been developed for the case where there is one-to-one correspondence between the data of the two modalities. In this paper, we propose a simple, yet effective generalized hashing framework which can work for all the different scenarios, while preserving the semantic distance between the data points. The approach first learns the optimum hash codes for the two modalities simultaneously, so as to preserve the semantic similarity between the data points, and then learns the hash functions to map from the features to the hash codes. Extensive experiments on single label dataset like Wiki and multi-label datasets like NUS-WIDE, Pascal and LabelMe under all the different scenarios and comparisons with the state-of-the-art shows the effectiveness of the proposed approach.
**********

FC4: Fully Convolutional Color Constancy With Confidence-Weighted Pooling
Yuanming Hu, Baoyuan Wang, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4085-4094
Improvements in color constancy have arisen from the use of convolutional neural networks (CNNs). However, the patch-based CNNs that exist for this problem are faced with the issue of estimation ambiguity, where a patch may contain insufficient information to establish a unique or even a limited possible range of illumination colors. Image patches with estimation ambiguity not only appear with great frequency in photographs, but also significantly degrade the quality of network training and inference. To overcome this problem, we present a fully convolutional network architecture in which patches throughout an image can carry different confidence weights according to the value they provide for color constancy e

stimation. These confidence weights are learned and applied within a novel pooli
ng layer where the local estimates are merged into a global solution. With this
formulation, the network is able to determine "what to learn" and "how to pool"
automatically from color constancy datasets without additional supervision. The
proposed network also allows for end-to-end training, and achieves higher effici
ency and accuracy. On standard benchmarks, our network outperforms the previous
state-of-the-art while achieving 120x greater efficiency.
**********

Template-Based Monocular 3D Recovery of Elastic Shapes Using Lagrangian Multipli
ers

Nazim Haouchine, Stephane Cotin; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2017, pp. 4095-4103

We present in this paper an efficient template-based method for 3D recovery of e
lastic shapes from a fixed monocular camera. By exploiting the object's elastici
ty, in contrast to isometric methods that use inextensibility constraints, a lar
ge range of deformations can be handled. Our method is expressed as a saddle poi
nt problem using Lagrangian multipliers resulting in a linear system which unifi
es both mechanical and optical constraints and integrates Dirichlet boundary con
ditions, whether they are fixed or free. We experimentally show that no prior kn
owledge on material properties is needed, which exhibit the generic usability of
 our method with elastic and inelastic objects with different kinds of materials
. Comparisons with existing techniques are conducted on synthetic and real elast
ic objects with strains ranging from 25% to 130% resulting to low errors.
**********

Discriminative Optimization: Theory and Applications to Point Cloud Registration

Jayakorn Vongkulbhisal, Fernando De la Torre, Joao P. Costeira; Proceedings of t
he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.
4104-4112

Many computer vision problems are formulated as the optimization of a cost funct
ion. This approach faces two main challenges: (1) designing a cost function with
 a local optimum at an acceptable solution, and (2) developing an efficient nume
rical method to search for one (or multiple) of these local optima. While design
ing such functions is feasible in the noiseless case, the stability and location
 of local optima are mostly unknown under noise, occlusion, or missing data. In
practice, this can result in undesirable local optima or not having a local opti
mum in the expected place. On the other hand, numerical optimization algorithms
in high-dimensional spaces are typically local and often rely on expensive first
 or second order information to guide the search. To overcome these limitations,
 this paper proposes Discriminative Optimization (DO), a method that learns sear
ch directions from data without the need of a cost function. Specifically, DO ex
plicitly learns a sequence of updates in the search space that leads to stationa
ry points that correspond to desired solutions. We provide a formal analysis of
DO and illustrate its benefits in the problem of 2D and 3D point cloud registrat
ion both in synthetic and range-scan data. We show that DO outperforms state-of-
the-art algorithms by a large margin in terms of accuracy, robustness to perturb
ations, and computational efficiency.
**********

Fine-Grained Recognition of Thousands of Object Categories With Single-Example T
raining

Leonid Karlinsky, Joseph Shtok, Yochay Tzur, Asaf Tzadok; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4113-4
122

We approach the problem of fast detection and recognition of a large number (tho
usands) of object categories while training on a very limited amount of examples
, usually one per category. Examples of this task include: (i) detection of reta
il products, where we have only one studio image of each product available for t
raining; (ii) detection of brand logos; and (iii) detection of 3D objects and th
eir respective poses within a static 2D image, where only a sparse subset of (pa
rtial) object views is available for training, with a single example for each vi
ew. Building a detector based on so few examples presents a significant challeng

e for the current top-performing (deep) learning based techniques, which require large amounts of data to train. Our approach for this task is based on a non-parametric probabilistic model for initial detection, CNN-based refinement and temporal integration where applicable. We successfully demonstrate its usefulness in a variety of experiments on both existing and our own benchmarks achieving state-of-the-art performance.
**********

Deep Co-Occurrence Feature Learning for Visual Object Recognition
Ya-Fang Shih, Yang-Ming Yeh, Yen-Yu Lin, Ming-Fang Weng, Yi-Chang Lu, Yung-Yu Chuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4123-4132
This paper addresses three issues in integrating part-based representations into convolutional neural networks (CNNs) for object recognition. First, most part-based models rely on a few pre-specified object parts. However, the optimal object parts for recognition often vary from category to category. Second, acquiring training data with part-level annotation is labor-intensive. Third, modeling spatial relationships between parts in CNNs often involves an exhaustive search of part templates over multiple network streams. We tackle the three issues by introducing a new network layer, called co-occurrence layer. It can extend a convolutional layer to encode the co-occurrence between the visual parts detected by the numerous neurons, instead of a few pre-specified parts. To this end, the feature maps serve as both filters and images, and mutual correlation filtering is conducted between them. The co-occurrence layer is end-to-end trainable. The resultant co-occurrence features are rotation- and translation-invariant, and are robust to object deformation. By applying this new layer to the VGG-16 and ResNet-152, we achieve the recognition rates of 83.6% and 85.8% on the Caltech-UCSD bird benchmark, respectively. The source code is available at https://github.com/yafangshih/Deep-COOC.
**********

A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning
Junho Yim, Donggyu Joo, Jihoon Bae, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4133-4141
We introduce a novel technique for knowledge transfer, where knowledge from a pretrained deep neural network (DNN) is distilled and transferred to another DNN. As the DNN performs a mapping from the input space to the output space through many layers sequentially, we define the distilled knowledge to be transferred in terms of flow between layers, which is calculated by computing the inner product between features from two layers. When we compare the student DNN and the original network with the same size as the student DNN but trained without a teacher network, the proposed method of transferring the distilled knowledge as the flow between two layers exhibits three important phenomena : (1) the student DNN that learns the distilled knowledge is optimized much faster than the original model; (2) the student DNN outperforms the original DNN; and (3) the student DNN can learn the distilled knowledge from a teacher DNN that is trained at a different task, and the student DNN outperforms the original DNN that is trained from scratch.
**********

What Is and What Is Not a Salient Object? Learning Salient Object Detector by Ensembling Linear Exemplar Regressors
Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, Yu Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4142-4150
Finding what is and what is not a salient object can be helpful in developing better features and models in salient object detection (SOD). In this paper, we investigate the images that are selected and discarded in constructing a new SOD dataset and find that many similar candidates, complex shape and low objectness are three main attributes of many non-salient objects. Moreover, objects may have diversified attributes that make them salient. As a result, we propose a novel salient object detector by ensembling linear exemplar regressors. We first selec

t reliable foreground and background seeds using the boundary prior and then ado
pt locally linear embedding (LLE) to conduct manifold-preserving foregroundness
propagation. In this manner, a foregroundness map can be generated to roughly po
p-out salient objects and suppress non-salient ones with many similar candidates
. Moreover, we extract the shape, foregroundness and attention descriptors to ch
aracterize the extracted object proposals, and a linear exemplar regressor is tr
ained to encode how to detect salient proposals in a specific image. Finally, va
rious linear exemplar regressors are ensembled to form a single detector that ad
apts to various scenarios. Extensive experimental results on 5 dataset and the n
ew SOD dataset show that our approach outperforms 9 state-of-art methods.
**********

Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes
Tobias Pohlen, Alexander Hermans, Markus Mathias, Bastian Leibe; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.
4151-4160
Semantic image segmentation is an essential component of modern autonomous drivi
ng systems, as an accurate understanding of the surrounding scene is crucial to
navigation and action planning. Current state-of-the-art approaches in semantic
image segmentation rely on pre-trained networks that were initially developed fo
r classifying images as a whole. While these networks exhibit outstanding recogn
ition performance (i.e., what is visible?), they lack localization accuracy (i.e
., where precisely is something located?). Therefore, additional processing step
s have to be performed in order to obtain pixel-accurate segmentation masks at t
he full image resolution. To alleviate this problem we propose a novel ResNet-li
ke architecture that exhibits strong localization and recognition performance. W
e combine multi-scale context with pixel-level accuracy by using two processing
streams within our network: One stream carries information at the full image res
olution, enabling precise adherence to segment boundaries. The other stream unde
rgoes a sequence of pooling operations to obtain robust features for recognition
. The two streams are coupled at the full image resolution using residuals. With
out additional processing steps and without pre-training, our approach achieves
an intersection-over-union score of 71.8% on the Cityscapes dataset.
**********

Optical Flow Estimation Using a Spatial Pyramid Network
Anurag Ranjan, Michael J. Black; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2017, pp. 4161-4170
We learn to compute optical flow by combining a classical spatial-pyramid formul
ation with deep learning. This estimates large motions in a coarse-to-fine appro
ach by warping one image of a pair at each pyramid level by the current flow est
imate and computing an update to the flow. Instead of the standard minimization
of an objective function at each pyramid level, we train one deep network per le
vel to compute the flow update. Unlike the recent FlowNet approach, the networks
do not need to deal with large motions; these are dealt with by the pyramid. Th
is has several advantages. First, our Spatial Pyramid Network (SPyNet) is much s
impler and 96% smaller than FlowNet in terms of model parameters. This makes it
more efficient and appropriate for embedded applications. Second, since the flow
at each pyramid level is small ( < 1 pixel), a convolutional approach applied t
o pairs of warped images is appropriate. Third, unlike FlowNet, the learned conv
olution filters appear similar to classical spatio-temporal filters, giving insi
ght into the method and how to improve it. Our results are more accurate than Fl
owNet on most standard benchmarks, suggesting a new direction of combining class
ical flow methods with deep learning.
**********

Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action
Recognition
Junwu Weng, Chaoqun Weng, Junsong Yuan; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2017, pp. 4171-4180
Motivated by previous success of using non-parametric methods to recognize objec
ts, e.g., NBNN, we extend it to recognize actions using skeletons. Each 3D actio
n is presented by a sequence of 3D poses. Similar to NBNN, our proposed Spatio-T

emporal-NBNN applies stage-to-class distance to classify actions. However, ST-NB
NN takes the spatio-temporal structure of 3D actions into consideration and rela
xes the Naive Bayes assumption of NBNN. Specifically, ST-NBNN adopts bilinear cl
assifiers to identify both key temporal stages as well as spatial joints for act
ion classification. Although only using a linear classifier, experiments on thre
e benchmark datasets show that by combining the strength of both non-parametric
 and parametric models, ST-NBNN can achieve competitive performance compared with
 state-of-the-art results using sophisticated models such as deep learning. More
over, by identifying key skeleton joints and temporal stages for each action cla
ss, our ST-NBNN can capture the essential spatio-temporal patterns that play key
 roles of recognizing actions, which is not always achievable by using end-to-en
d models.
**********

GMS: Grid-based Motion Statistics for Fast, Ultra-Robust Feature Correspondence
JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, M
ing-Ming Cheng; Proceedings of the IEEE Conference on Computer Vision and Patter
n Recognition (CVPR), 2017, pp. 4181-4190

Incorporating smoothness constraints into feature matching is known to enable ul
tra-robust matching. However, such formulations are both complex and slow, makin
g them unsuitable for video applications. This paper proposes GMS (Grid-based Mo
tion Statistics), a simple means of encapsulating motion smoothness as the stati
stical likelihood of a certain number of matches in a region. GMS enables transl
ation of high match numbers into high match quality. This provides a real-time,
ultra-robust correspondence system. Evaluation on videos, with low textures, blu
rs and wide-baselines show GMS consistently out-performs other real-time matcher
s and can achieve parity with more sophisticated, much slower techniques.
**********

Detailed, Accurate, Human Shape Estimation From Clothed 3D Scan Sequences
Chao Zhang, Sergi Pujades, Michael J. Black, Gerard Pons-Moll; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4
191-4200

We address the problem of estimating human pose and body shape from 3D scans ove
r time. Reliable estimation of 3D body shape is necessary for many applications
including virtual try-on, health monitoring, and avatar creation for virtual rea
lity. Scanning bodies in minimal clothing, however, presents a practical barrier
 to these applications. We address this problem by estimating body shape under c
lothing from a sequence of 3D scans. Previous methods that have exploited body m
odels produce smooth shapes lacking personalized details. We contribute a new ap
proach to recover a personalized shape of the person. The estimated shape deviat
es from a parametric model to fit the 3D scans. We demonstrate the method using
high quality 4D data as well as sequences of visual hulls extracted from multi-v
iew images.  We also make available BUFF, a new 4D dataset that enables quantita
tive evaluation (http://buff.is.tue.mpg.de). Our method outperforms the state of
 the art  in both pose estimation and shape estimation, qualitatively and quanti
tatively.
**********

Active Convolution: Learning the Shape of Convolution for Image Classification
Yunho Jeon, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and
 Pattern Recognition (CVPR), 2017, pp. 4201-4209

In recent years, deep learning has achieved great success in many computer visio
n applications. Convolutional neural networks (CNNs) have lately emerged as a ma
jor approach to image classification. Most research on CNNs thus far has focused
 on developing architectures such as the Inception and residual networks. The co
nvolution layer is the core of the CNN, but few studies have addressed the convo
lution unit itself. In this paper, we introduce a convolution unit called the ac
tive convolution unit (ACU). A new convolution has no fixed shape, because of wh
ich we can define any form of convolution. Its shape can be learned through back
propagation during training. Our proposed unit has a few advantages. First, the
ACU is a generalization of convolution; it can define not only all conventional
convolutions, but also convolutions with fractional pixel coordinates. We can fr

eely change the shape of the convolution, which provides greater freedom to form CNN structures. Second, the shape of the convolution is learned while training and there is no need to tune it by hand. Third, the ACU can learn better than a conventional unit, where we obtained the improvement simply by changing the conventional convolution to an ACU. We tested our proposed method on plain and residual networks, and the results showed significant improvement using our method on various datasets and architectures in comparison with the baseline. Code is available at https://github.com/jyh2986/Active-Convolution.

**********

Video Desnowing and Deraining Based on Matrix Decomposition

Weihong Ren, Jiandong Tian, Zhi Han, Antoni Chan, Yandong Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4210-4219

The existing snow/rain removal methods often fail for heavy snow/rain and dynamic scene. One reason for the failure is due to the assumption that all the snowflakes/rain streaks are sparse in snow/rain scenes. The other is that the existing methods often can not differentiate moving objects and snowflakes/rain streaks. In this paper, we propose a model based on matrix decomposition for video desnowing and deraining to solve the problems mentioned above. We divide snowflakes/rain streaks into two categories: sparse ones and dense ones. With background fluctuations and optical flow information, the detection of moving objects and sparse snowflakes/rain streaks is formulated as a multi-label Markov Random Fields (MRFs). As for dense snowflakes/rain streaks, they are considered to obey Gaussian distribution. The snowflakes/rain streaks, including sparse ones and dense ones, in scene backgrounds are removed by low-rank representation of the backgrounds. Meanwhile, a group sparsity term in our model is designed to filter snow/rain pixels within the moving objects. Experimental results show that our proposed model performs better than the state-of-the-art methods for snow and rain removal.

**********

Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos

Jie Song, Limin Wang, Luc Van Gool, Otmar Hilliges; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4220-4229

Deep ConvNets have been shown to be effective for the task of human pose estimation from single images. However, several challenging issues arise in the video-based case such as self-occlusion, motion blur, and uncommon poses with few or no examples in the training data. Temporal information can provide additional cues about the location of body joints and help to alleviate these issues. In this paper, we propose a deep structured model to estimate a sequence of human poses in unconstrained videos. This model can be efficiently trained in an end-to-end manner and is capable of representing the appearance of body joints and their spatio-temporal relationships simultaneously. Domain knowledge about the human body is explicitly incorporated into the network providing effective priors to regularize the skeletal structure and to enforce temporal consistency. The proposed end-to-end architecture is evaluated on two widely used benchmarks for video-based pose estimation (Penn Action and JHMDB datasets). Our approach outperforms several state-of-the-art methods.

**********

Self-Supervised Learning of Visual Features Through Embedding Images Into Text Topic Spaces

Lluis Gomez, Yash Patel, Marcal Rusinol, Dimosthenis Karatzas, C. V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4230-4239

End-to-end training from scratch of current deep architectures for new computer vision problems would require Imagenet-scale datasets, and this is not always possible. In this paper we present a method that is able to take advantage of freely available multi-modal content to train computer vision algorithms without human supervision. We put forward the idea of performing self-supervised learning of visual features by mining a large scale corpus of multi-modal (text and image) documents. We show that discriminative visual features can be learnt efficientl

y by training a CNN to predict the semantic context in which a particular image is more probable to appear as an illustration. For this we leverage the hidden s emantic structures discovered in the text corpus with a well-known topic modelin g technique. Our experiments demonstrate state of the art performance in image c lassification, object detection, and multi-modal retrieval compared to recent se lf-supervised or natural-supervised approaches.
**********

Coarse-To-Fine Segmentation With Shape-Tailored Continuum Scale Spaces
Naeemullah Khan, Byung-Woo Hong, Anthony Yezzi, Ganesh Sundaramoorthi; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 7, pp. 4240-4249
We formulate an energy for segmentation that is designed to have preference for segmenting the coarse over fine structure of the image, without smoothing across boundaries of regions. The energy is formulated by integrating a continuum of s cales from a scale space computed from the heat equation within regions. We show that the energy can be optimized without computing a continuum of scales, but i nstead from a single scale. This makes the method computationally efficient in c omparison to energies using a discrete set of scales. We apply our method to tex ture and motion segmentation. Experiments on benchmark datasets show that a cont inuum of scales leads to better segmentation accuracy over discrete scales and o ther competing methods.
**********

Minimum Delay Moving Object Detection
Dong Lao, Ganesh Sundaramoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4250-4259
We present a general framework and method for detection of an object in a video based on apparent motion. The object moves relative to background motion at some unknown time in the video, and the goal is to detect and segment the object as soon it moves in an online manner. Due to unreliability of motion between frames , more than two frames are needed to reliably detect the object. Our method is d esigned to detect the object(s) with minimum delay, i.e., frames after the objec t moves, constraining the false alarms. Experiments on a new extensive dataset f or moving object detection show that our method achieves less delay for all fals e alarm constraints than existing state-of-the-art.
**********

Hyper-Laplacian Regularized Unidirectional Low-Rank Tensor Recovery for Multispe ctral Image Denoising
Yi Chang, Luxin Yan, Sheng Zhong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4260-4268
Recent low-rank based matrix/tensor recovery methods have been widely explored i n multispectral images (MSI) denoising. These methods, however, ignore the diffe rence of the intrinsic structure correlation along spatial sparsity, spectral co rrelation and non-local self-similarity mode. In this paper, we go further by gi ving a detailed analysis about the rank properties both in matrix and tensor cas es, and figure out the non-local self-similarity is the key ingredient, while th e low-rank assumption of others may not hold. This motivates us to design a simp le yet effective unidirectional low-rank tensor recovery model that is capable o f truthfully capturing the intrinsic structure correlation with reduced computat ional burden. However, the low-rank models suffer from the ringing artifacts, du e to the aggregation of overlapped patches/cubics. While previous methods resort to spatial information, we offer a new perspective by utilizing the exclusively spectral information in MSIs to address the issue. The analysis-based hyper-Lap lacian prior is introduced to model the global spectral structures, so as to ind irectly alleviate the ringing artifacts in spatial domain. The advantages of the proposed method over the existing ones are multi-fold: more reasonably structur e correlation representability, less processing time, and less artifacts in the overlapped regions. The proposed method is extensively evaluated on several benc hmarks, and significantly outperforms state-of-the-art MSI denoising methods.
**********

Online Asymmetric Similarity Learning for Cross-Modal Retrieval

Yiling Wu, Shuhui Wang, Qingming Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4269-4278
Cross-modal retrieval has attracted intensive attention in recent years. Measuring the semantic similarity between heterogeneous data objects is an essential yet challenging problem in cross-modal retrieval. In this paper, we propose an online learning method to learn the similarity function between heterogeneous modalities by preserving the relative similarity in the training data, which is modeled as a set of bi-directional hinge loss constraints on the cross-modal training triplets. The overall online similarity function learning problem is optimized by the margin based Passive-Aggressive algorithm. We further extend the approach to learn similarity function in reproducing kernel Hilbert spaces by kernelizing the approach and combining multiple kernels derived from different layers of the CNN features using the Hedging algorithm. Theoretical mistake bounds are given for our methods. Experiments conducted on real world datasets well demonstrate the effectiveness of our methods.
**********

Latent Multi-View Subspace Clustering
Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, Xiaochun Cao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4279-4287
In this paper, we propose a novel Latent Multi-view Subspace Clustering (LMSC) method, which clusters data points with latent representation and simultaneously explores underlying complementary information from multiple views. Unlike most existing single view subspace clustering methods that reconstruct data points using original features, our method seeks the underlying latent representation and simultaneously performs data reconstruction based on the learned latent representation. With the complementarity of multiple views, the latent representation could depict data themselves more comprehensively than each single view individually, accordingly makes subspace representation more accurate and robust as well. The proposed method is intuitive and can be optimized efficiently by using the Augmented Lagrangian Multiplier with Alternating Direction Minimization (ALM-ADM) algorithm. Extensive experiments on benchmark datasets have validated the effectiveness of our proposed method.
**********

Unsupervised Vanishing Point Detection and Camera Calibration From a Single Manhattan Image With Radial Distortion
Michel Antunes, Joao P. Barreto, Djamila Aouada, Bjorn Ottersten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4288-4296
The article concerns the automatic calibration of a camera with radial distortion from a single image. It is known that, under the mild assumption of square pixels and zero skew, lines in the scene project into circles in the image, and three lines suffice to calibrate the camera up to an ambiguity between focal length and radial distortion. The calibration results highly depend on accurate circle estimation, which is hard to accomplish because lines tend to project into short circular arcs. To overcome this problem, we show that, given a short circular arc edge, it is possible to robustly determine a line that goes through the center of the corresponding circle. These lines, henceforth called Lines of Circle Centres (LCCs), are used in a new method that detects sets of parallel lines and estimates the calibration parameters, including the center and amount of distortion, focal length, and camera orientation with respect to the Manhattan frame. Extensive experiments in both semi-synthetic and real images show that our algorithm outperforms state-of-the-art approaches in unsupervised calibration from a single image, while providing more information.
**********

Re-Sign: Re-Aligned End-To-End Sequence Modelling With Deep Recurrent CNN-HMMs
Oscar Koller, Sepehr Zargaran, Hermann Ney; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4297-4305
This work presents an iterative re-alignment approach applicable to visual sequence labelling tasks such as gesture recognition, activity recognition and contin

uous sign language recognition. Previous methods dealing with video data usuall y rely on given frame labels to train their classifiers. However, looking at rec ent data sets, these labels often tend to be noisy which is commonly overseen. We propose an algorithm that treats the provided training labels as weak labels and refines the label-to-image alignment on-the-fly in a weakly supervised fashi on. Given a series of frames and sequence-level labels, a deep recurrent CNN-B LSTM network is trained end-to-end. Embedded into an HMM the resulting deep mode l corrects the frame labels and continuously improves its performance in several re-alignments. We evaluate on two challenging publicly available sign recogni tion benchmark data sets featuring over 1000 classes. We outperform the state-of -the-art by up to 10% absolute and 30% relative.

**********

Improving Interpretability of Deep Neural Networks With Semantic Information

Yinpeng Dong, Hang Su, Jun Zhu, Bo Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4306-4314

Interpretability of deep neural networks (DNNs) is essential since it enables us ers to understand the overall strengths and weaknesses of the models, conveys an understanding of how the models will behave in the future, and how to diagnose and correct potential problems. However, it is challenging to reason about what a DNN actually does due to its opaque or black-box nature. To address this issue , we propose a novel technique to improve the interpretability of DNNs by levera ging the rich semantic information embedded in human descriptions. By concentrat ing on the video captioning task, we first extract a set of semantically meaning ful topics from the human descriptions that cover a wide range of visual concept s, and integrate them into the model with an interpretive loss. We then propose a prediction difference maximization algorithm to interpret the learned features of each neuron. Experimental results demonstrate its effectiveness in video cap tioning using the interpretable features, which can also be transferred to video action recognition. By clearly understanding the learned features, users can ea sily revise false predictions via a human-in-the-loop procedure.

**********

Social Scene Understanding: End-To-End Multi-Person Action Localization and Coll ective Activity Recognition

Timur Bagautdinov, Alexandre Alahi, Francois Fleuret, Pascal Fua, Silvio Savares e; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4315-4324

We present a unified framework for understanding human social behaviors in raw i mage sequences. Our model jointly detects multiple individuals, infers their soc ial actions, and estimates the collective actions with a single feed-forward pas s through a neural network. We propose a single architecture that does not rely on external detection algorithms but rather is trained end-to-end to generate de nse proposal maps that are refined via a novel inference scheme. The temporal co nsistency is handled via a person-level matching Recurrent Neural Network. The c omplete model takes as input a sequence of frames and outputs detections along w ith the estimates of individual actions and collective activities. We demonstrat e state-of-the-art performance of our algorithm on multiple publicly available b enchmarks.

**********

UntrimmedNets for Weakly Supervised Action Recognition and Detection

Limin Wang, Yuanjun Xiong, Dahua Lin, Luc Van Gool; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4325-4334

Current action recognition methods heavily rely on trimmed videos for model trai ning. However, it is expensive and time-consuming to acquire a large-scale trimm ed video dataset. This paper presents a new weakly supervised architecture, call ed UntrimmedNet, which is able to directly learn action recognition models from untrimmed videos without the requirement of temporal annotations of action insta nces. Our UntrimmedNet couples two important components, the classification modu le and the selection module, to learn the action models and reason about the tem poral duration of action instances, respectively. These two components are imple mented with feed-forward networks, and UntrimmedNet is therefore an end-to-end t

rainable architecture. We exploit the learned models for action recognition (WSR ) and detection (WSD) on the untrimmed video datasets of THUMOS14 and ActivityNe t. Although our UntrimmedNet only employs weak supervision, our method achieves performance superior or comparable to that of those strongly supervised approach es on these two datasets.

**********

Multi-Task Correlation Particle Filter for Robust Object Tracking

Tianzhu Zhang, Changsheng Xu, Ming-Hsuan Yang; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4335-4343

In this paper, we propose a multi-task correlation particle filter (MCPF) for ro bust visual tracking. We first present the multi-task correlation filter (MCF) t hat takes the interdependencies among different features into account to learn c orrelation filters jointly. The proposed MCPF is designed to exploit and complem ent the strength of a MCF and a particle filter. Compared with existing tracking methods based on correlation filters and particle filters, the proposed tracker has several advantages. First, it can shepherd the sampled particles toward the modes of the target state distribution via the MCF, thereby resulting in robust tracking performance. Second, it can effectively handle large-scale variation v ia a particle sampling strategy. Third, it can effectively maintain multiple mod es in the posterior density using fewer particles than conventional particle fil ters, thereby lowering the computational cost. Extensive experimental results on three benchmark datasets demonstrate that the proposed MCPF performs favorably against the state-of-the-art methods.

**********

Improving Training of Deep Neural Networks via Singular Value Bounding

Kui Jia, Dacheng Tao, Shenghua Gao, Xiangmin Xu; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4344-4352

Deep learning methods achieve great success recently on many computer vision pro blems. In spite of these practical successes, optimization of deep networks rema ins an active topic in deep learning research. In this work, we focus on investi gation of the network solution properties that can potentially lead to good perf ormance. Our research is inspired by theoretical and empirical results that use orthogonal matrices to initialize networks, but we are interested in investigati ng how orthogonal weight matrices perform when network training converges. To th is end, we propose to constrain the solutions of weight matrices in the orthogon al feasible set during the whole process of network training, and achieve this b y a simple yet effective method called Singular Value Bounding (SVB). In SVB, al l singular values of each weight matrix are simply bounded in a narrow band arou nd the value of 1. Based on the same motivation, we also propose Bounded Batch N ormalization (BBN), which improves Batch Normalization by removing its potential risk of ill-conditioned layer transform. We present both theoretical and empiri cal results to justify our proposed methods. Experiments on benchmark image clas sification datasets show the efficacy of our proposed SVB and BBN. In particular , we achieve the state-of-the-art results of 3.06% error rate on CIFAR10 and 16. 90% on CIFAR100, using off-the-shelf network architectures (Wide ResNets). Our p reliminary results on ImageNet also show the promise in large-scale learning. We release the implementation code of our methods at www.aperture-lab.net/research /svb.

**********

Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Ne twork

Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, Jian Sun; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4353-4 361

Convolution Neural Network (CNN) has boosted the per- formanceofalotofcomputervi siontasks, likeimageclassi- fication [31], segmentation [25], and detection [28] . Based on the observations from [31, 32, 14], recent model design- ers prefer t o employ stacking of small kernels, like 3 x 3 over large-size filters. However, in the field of semantic seg- mentation, where we need to perform dense per-pix el pre- diction, we find that large kernel plays an important role to relieve th

e contradictories when optimizing the classi- fication and localization tasks si multaneously. Following the design principle of large-size kernel, We propose th e Global Convolutional Network to address both the classi- fication and localiza tion issue in the semantic segmentation task. To further refine the object categ ory boundaries, we presentBoundaryRefinementblockbasedonresidualstruc- ture. Qua litatively, our model achieves state-of-art perfor- mance on two public benchmar ks and outperforms previous results on a large margin, 82.2% (vs 80.2%) on PASCA L VOC 2012 dataset and 76.9% (vs 71.8%) on Cityscapes dataset.
**********

Neural Aggregation Network for Video Face Recognition
Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, Gan g Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni tion (CVPR), 2017, pp. 4362-4371
This paper presents a Neural Aggregation Network (NAN) for video face recognitio n. The network takes a face video or face image set of a person with a variable number of face images as its input, and produces a compact, fixed-dimension feat ure representation for recognition. The whole network is composed of two modules . The feature embedding module is a deep Convolutional Neural Network (CNN) whic h maps each face image to a feature vector. The aggregation module consists of t wo attention blocks which adaptively aggregate the feature vectors to form a sin gle feature inside the convex hull spanned by them. Due to the attention mechani sm, the aggregation is invariant to the image order. Our NAN is trained with a s tandard classification or verification loss without any extra supervision signal , and we found that it automatically learns to advocate high-quality face images while repelling low-quality ones such as blurred, occluded and improperly expos ed faces. The experiments on IJB-A, YouTube Face, Celebrity-1000 video face reco gnition benchmarks show that it consistently outperforms naive aggregation metho ds and achieves the state-of-the-art accuracy.
**********

Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Netwo rks
Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, Jiashi Feng; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4372-4381
We introduce a new problem of gaze anticipation on egocentric videos. This subst antially extends the conventional gaze prediction problem to future frames by no longer confining it on the current frame. To solve this problem, we propose a n ew generative adversarial neural network based model, Deep Future Gaze (DFG). DF G generates multiple future frames conditioned on the single current frame and a nticipates corresponding future gazes in next few seconds. It consists of two ne tworks: generator and discriminator. The generator uses a two-stream spatial tem poral convolution architecture (3D-CNN) explicitly untangling the foreground and the background to generate future frames. It then attaches another 3D-CNN for g aze anticipation based on these synthetic frames. The discriminator plays agains t the generator by differentiating the synthetic frames of the generator from th e real frames. Through competition with discriminator, the generator progressive ly improves quality of the future frames and thus anticipates future gaze better . Experimental results on the publicly available egocentric datasets show that D FG significantly outperforms all well-established baselines. Moreover, we demons trate that DFG achieves better performance of gaze prediction on current frames than state-of-the-art methods. This is due to benefiting from learning motion di scriminative representations in frame generation. We further contribute a new eg ocentric dataset (OST) in the object search task. DFG also achieves the best per formance for this challenging dataset.
**********

Simultaneous Stereo Video Deblurring and Scene Flow Estimation
Liyuan Pan, Yuchao Dai, Miaomiao Liu, Fatih Porikli; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4382-4391
Videos for outdoor scene often show unpleasant blur effects due to the large rel ative motion between the camera and the dynamic objects and large depth variatio

ns. Existing works typically focus monocular video deblurring. In this paper, we propose a novel approach to deblurring from stereo videos. In particular, we exploit the piece-wise planar assumption about the scene and leverage the scene flow information to deblur the image. Unlike the existing approach [31] which used a pre-computed scene flow, we propose a single framework to jointly estimate the scene flow and deblur the image, where the motion cues from scene flow estimation and blur information could reinforce each other, and produce superior results than the conventional scene flow estimation or stereo deblurring methods. We evaluate our method extensively on two available datasets and achieve significant improvement in flow estimation and removing the blur effect over the state-of-the-art methods.

**********

An Empirical Evaluation of Visual Question Answering for Novel Objects

Santhosh K. Ramakrishnan, Ambar Pal, Gaurav Sharma, Anurag Mittal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4392-4401

We study the problem of answering questions about images in the harder setting, where the test questions and corresponding images contain novel objects, which were not queried about in the training data. Such setting is inevitable in real world--owing to the heavy tailed distribution of the visual categories, there would be some objects which would not be annotated in the train set. We show that the performance of two popular existing methods drop significantly (21-28%) when evaluated on novel objects cf. known objects. We propose methods which use large existing external corpora of (i) unlabeled text, i.e. books, and (ii) images tagged with classes, to achieve novel object based visual question answering. We systematically study both, an oracle case where the novel objects are known textually, as well as a fully automatic case without any explicit knowledge of the novel objects, but with the minimal assumption that the novel objects are semantically related to the existing objects in training. The proposed methods for novel object based visual question answering are modular and can potentially be used with many visual question answering architectures. We show consistent improvements with the two popular architectures and give qualitative analysis of the cases where the model does well and of those where it fails to bring improvements.

**********

Binary Constraint Preserving Graph Matching

Bo Jiang, Jin Tang, Chris Ding, Bin Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4402-4409

Graph matching is a fundamental problem in computer vision and pattern recognition area. In general, it can be formulated as an Integer Quadratic Programming (IQP) problem. Since it is NP-hard, approximate relaxations are required. In this paper, a new graph matching method has been proposed. There are three main contributions of the proposed method: (1) we propose a new graph matching relaxation model, called Binary Constraint Preserving Graph Matching (BPGM), which aims to incorporate the discrete binary mapping constraints more in graph matching relaxation. Our BPGM is motivated by a new observation that the discrete binary constraints in IQP matching problem can be represented (or encoded) exactly by a l2-norm constraint. (2) An effective projection algorithm has been derived to solve BPGM model. (3) Using BPGM, we propose a path-following strategy to optimize IQP matching problem and thus obtain a desired discrete solution at convergence. Promising experimental results show the effectiveness of the proposed method.

**********

Exploiting Saliency for Object Segmentation From Image Level Labels

Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4410-4419

There have been remarkable improvements in the semantic labelling task in the recent years. However, the state of the art methods rely on large-scale pixel-level annotations. This paper studies the problem of training a pixel-wise semantic labeller network from image-level annotations of the present object classes. Recently, it has been shown that high quality seeds indicating discriminative objec

t regions can be obtained from image-level labels. Without additional informatio n, obtaining the full extent of the object is an inherently ill-posed problem du e to co-occurrences. We propose using a saliency model as additional information and hereby exploit prior knowledge on the object extent and image statistics. W e show how to combine both information sources in order to recover 80% of the fu lly supervised performance - which is the new state of the art in weakly supervi sed training for pixel-wise semantic labelling.

**********

Predicting Salient Face in Multiple-Face Videos

Yufan Liu, Songyang Zhang, Mai Xu, Xuming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4420-4428

Although the recent success of convolutional neural network (CNN) advances state -of-the-art saliency prediction in static images, few work has addressed the pro blem of predicting attention in videos. On the other hand, we find that the atte ntion of different subjects consistently focuses on a single face in each frame of videos involving multiple faces. Therefore, we propose in this paper a novel deep learning (DL) based method to predict salient face in multiple-face videos, which is capable of learning features and transition of salient faces across v ideo frames. In particular, we first learn a CNN for each frame to locate salien t face. Taking CNN features as input, we develop a multiple-stream long short-te rm memory (M-LSTM) network to predict the temporal transition of salient faces i n video sequences. To evaluate our DL-based method, we build a new eye-tracking database of multiple-face videos. The experimental results show that our method outperforms the prior state-of-the-art methods in predicting visual attention o n faces in multiple-face videos.

**********

SPFTN: A Self-Paced Fine-Tuning Network for Segmenting Objects in Weakly Labelle d Videos

Dingwen Zhang, Le Yang, Deyu Meng, Dong Xu, Junwei Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4429-443 7

Object segmentation in weakly labelled videos is an interesting yet challenging task, which aims at learning to perform category-specific video object segmentat ion by only using video-level tags. Existing works in this research area might s till have some limitations, e.g., lack of effective DNN-based learning framework s, under-exploring the context information, and requiring to leverage the unstab le negative video collection, which prevent them from obtaining more promising p erformance. To this end, we propose a novel self-paced fine-tuning network (SPFT N)-based framework, which could learn to explore the context information within the video frames and capture adequate object semantics without using the negativ e videos. To perform weakly supervised learning based on the deep neural network , we make the earliest effort to integrate the self-paced learning regime and th e deep neural network into a unified and compatible framework, leading to the se lf-paced fine-tuning network. Comprehensive experiments on the large-scale YouTu be-Objects and DAVIS datasets demonstrate that the proposed approach achieves su perior performance as compared with other state-of-the-art methods as well as th e baseline networks and models.

**********

Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition

Jianlong Fu, Heliang Zheng, Tao Mei; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2017, pp. 4438-4446

Recognizing fine-grained categories (e.g., bird species) is difficult due to the challenges of discriminative region localization and fine-grained feature learn ing. Existing approaches predominantly solve these challenges independently, whi le neglecting the fact that region detection and fine-grained feature learning a re mutually correlated and thus can reinforce each other. In this paper, we prop ose a novel recurrent attention convolutional neural network (RA-CNN) which recu rsively learns discriminative region attention and region-based feature represen tation at multiple scales in a mutual reinforced way. The learning at each scale

consists of a classification sub-network and an attention proposal sub-network (APN). The APN starts from full images, and iteratively generates region attention from coarse to fine by taking previous prediction as a reference, while the finer scale network takes as input an amplified attended region from previous scale in a recurrent way. The proposed RA-CNN is optimized by an intra-scale classification loss and an inter-scale ranking loss, to mutually learn accurate region attention and fine-grained representation. RA-CNN does not need bounding box/part annotations and can be trained end-to-end. We conduct comprehensive experiments and show that RA-CNN achieves the best performance in three fine-grained tasks, with relative accuracy gains of 3.3%, 3.7%, 3.8%, on CUB Birds, Stanford Dogs and Stanford Cars, respectively.

**********

From Local to Global: Edge Profiles to Camera Motion in Blurred Images

Subeesh Vasu, A. N. Rajagopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4447-4456

In this work, we investigate the relation between the edge profiles present in a motion blurred image and the underlying camera motion responsible for causing the motion blur. While related works on camera motion estimation (CME) rely on the strong assumption of space-invariant blur, we handle the challenging case of general camera motion. We first show how edge profiles `alone' can be harnessed to perform direct CME from a single observation. While it is routine for conventional methods to jointly estimate the latent image too through alternating minimization, our above scheme is best-suited when such a pursuit is either impractical or inefficacious. For applications that actually favor an alternating minimization strategy, the edge profiles can serve as a valuable cue. We incorporate a suitably derived constraint from edge profiles into an existing blind deblurring framework and demonstrate improved restoration performance. Experiments reveal that this approach yields state-of-the-art results for the blind deblurring problem.

**********

On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation

Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Luigi Di Stefano, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4457-4466

Camera relocalisation is an important problem in computer vision, with applications in simultaneous localisation and mapping, virtual/augmented reality and navigation. Common techniques either match the current image against keyframes with known poses coming from a tracker, or establish 2D-to-3D correspondences between keypoints in the current image and points in the scene in order to estimate the camera pose. Recently, regression forests have become a popular alternative to establish such correspondences. They achieve accurate results, but must be trained offline on the target scene, preventing relocalisation in new environments. In this paper, we show how to circumvent this limitation by adapting a pre-trained forest to a new scene on the fly. Our adapted forests achieve relocalisation performance that is on par with that of offline forests, and our approach runs in under 150ms, making it desirable for real-time systems that require online relocalisation.

**********

Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, Jason Yosinski; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4467-4477

Generating high-resolution, photo-realistic images has been a long-standing goal in machine learning. Recently, Nguyen et al. 2016 showed one interesting way to synthesize novel images by performing gradient descent in the latent space of a generator network to maximize the activations of one or multiple neurons in a separate classifier network.  In this paper we extend this method by introducing an additional prior on the latent code, improving both sample quality and sample diversity, leading to a state-of-the-art generative model that produces high qu

ality images at higher resolutions (227x227) than previous generative models, and does so for all 1000 ImageNet categories. In addition, we provide a unified probabilistic interpretation of related activation maximization methods and call the general class of models "Plug and Play Generative Networks". PPGNs are composed of (1) a generator network G that is capable of drawing a wide range of image types and (2) a replaceable "condition" network C that tells the generator what to draw. We demonstrate generation of images conditioned on a class - when C is an ImageNet classification network - and also conditioned on a caption - when C is an image captioning network. Our method also improves the state of the art of Deep Multifaceted Feature Visualization, which involves synthetically generating the set of inputs that activate a neuron in order to better understand how deep neural networks operate. Finally, we show that our model performs reasonably well at the task of image inpainting. While we operate on images in this paper, the approach is modality agnostic and can be applied to many types of data.
**********

Domain Adaptation by Mixture of Alignments of Second- or Higher-Order Scatter Tensors
Piotr Koniusz, Yusuf Tas, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4478-4487
In this paper, we propose an approach to the domain adaptation, dubbed Second- or Higher-order Transfer of Knowledge (So-HoT), based on the mixture of alignments of second- or higher-order scatter statistics between the source and target domains. The human ability to learn from few labeled samples is a recurring motivation in the literature for domain adaptation. Towards this end, we investigate the supervised target scenario for which few labeled target training samples per category exist. Specifically, we utilize two CNN streams: the source and target networks fused at the classifier level. Features from the fully connected layers fc7 of each network are used to compute second- or even higher-order scatter tensors; one per network stream per class. As the source and target distributions are somewhat different despite being related, we align the scatters of the two network streams of the same class (within-class scatters) to a desired degree with our bespoke loss while maintaining good separation of the between-class scatters. We train the entire network in end-to-end fashion. We provide evaluations on the standard Office benchmark (visual domains) and RGB-D combined with Caltech256 (depth-to-rgb transfer). We attain state-of-the-art results.
**********

A Generative Model for Depth-Based Robust 3D Facial Pose Tracking
Lu Sheng, Jianfei Cai, Tat-Jen Cham, Vladimir Pavlovic, King Ngi Ngan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4488-4497
We consider the problem of depth-based robust 3D facial pose tracking under unconstrained scenarios with heavy occlusions and arbitrary facial expression variations. Unlike the previous depth-based discriminative or data-driven methods that require sophisticated training or manual intervention, we propose a generative framework that unifies pose tracking and face model adaptation on-the-fly. Particularly, we propose a statistical 3D face model that owns the flexibility to generate and predict the distribution and uncertainty underlying the face model. Moreover, unlike prior arts employing the ICP-based facial pose estimation, we propose a ray visibility constraint that regularizes the pose based on the face model's visibility against the input point cloud, which augments the robustness against the occlusions. The experimental results on Biwi and ICT-3DHP datasets reveal that the proposed framework is effective and outperforms the state-of-the-art depth-based methods.
**********

Single Image Reflection Suppression
Nikolaos Arvanitopoulos, Radhakrishna Achanta, Sabine Susstrunk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4498-4506
Reflections are a common artifact in images taken through glass windows. Automatically removing the reflection artifacts after the picture is taken is an ill-po

sed problem. Attempts to solve this problem using optimization schemes therefore rely on various prior assumptions from the physical world. Instead of removing reflections from a single image, which has met with limited success so far, we propose a novel approach to suppress reflections. It is based on a Laplacian data fidelity term and an l-zero gradient sparsity term imposed on the output. With experiments on artificial and real-world images we show that our reflection suppression method performs better than the state-of-the-art reflection removal techniques.

**********

Learning Non-Maximum Suppression

Jan Hosang, Rodrigo Benenson, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4507-4515

Object detectors have hugely profited from moving towards an end-to-end learning paradigm: proposals, fea tures, and the classifier becoming one neural network improved results two-fold on general object detection. One indispensable component is non-maximum suppression (NMS), a post-processing algorithm responsible for merging all detections that belong to the same object. The de facto standard NMS algorithm is still fully hand-crafted, suspiciously simple, and -- being based on greedy clustering with a fixed distance threshold -- forces a trade-off between recall and precision. We propose a new network architecture designed to perform NMS, using only boxes and their score. We report experiments for person detection on PETS and for general object categories on the COCO dataset. Our approach shows promise providing improved localization and occlusion handling.

**********

BRISKS: Binary Features for Spherical Images on a Geodesic Grid

Hao Guan, William A. P. Smith; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4516-4524

In this paper, we develop an interest point detector and binary feature descriptor for spherical images. We take as inspiration a recent framework developed for planar images, BRISK (Binary Robust Invariant Scalable Keypoints), and adapt the method to operate on spherical images. All of our processing is intrinsic to the sphere and avoids the distortion inherent in storing and indexing spherical images in a 2D representation. We discretise images on a spherical geodesic grid formed by recursive subdivision of a triangular mesh. This leads to a multiscale pixel grid comprising mainly hexagonal pixels that lends itself naturally to a spherical image pyramid representation. For interest point detection, we use a variant of the Accelerated Segment Test (AST) corner detector which operates on our geodesic grid. We estimate a continuous scale and location for features and descriptors are built by sampling onto a regular pattern in the tangent space. We evaluate repeatability, precision and recall on both synthetic spherical images with known ground truth correspondences and real images.

**********

Gaze Embeddings for Zero-Shot Image Classification

Nour Karessli, Zeynep Akata, Bernt Schiele, Andreas Bulling; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4525-4534

Zero-shot image classification using auxiliary information, such as attributes describing discriminative object properties, requires time-consuming annotation by domain experts. We instead propose a method that relies on human gaze as auxiliary information, exploiting that even non-expert users have a natural ability to judge class membership. We present a data collection paradigm that involves a discrimination task to increase the information content obtained from gaze data. Our method extracts discriminative descriptors from the data and learns a compatibility function between image and gaze using three novel gaze embeddings: Gaze Histograms (GH), Gaze Features with Grid (GFG) and Gaze Features with Sequence (GFS). We introduce two new gaze-annotated datasets for fine-grained image classification and show that human gaze data is indeed class discriminative, provides a competitive alternative to expert-annotated attributes, and outperforms other baselines for zero-shot image classification.

**********

A-Lamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural Network for Photo Aesthetic Assessment
Shuang Ma, Jing Liu, Chang Wen Chen; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2017, pp. 4535-4544
Deep convolutional neural networks (CNN) have recently been shown to generate pr omising results for aesthetics assessment. However, the performance of these dee p CNN methods is often compromised by the constraint that the neural network onl y takes the fixed-size input. To accommodate this requirement, input images need to be transformed via cropping, warping, or padding, which often alter image co mposition, reduce image resolution, or cause image distortion. Thus the aestheti cs of the original images is impaired because of potential loss of fine grained details and holistic image layout. However, such fine grained details and holist ic image layout is critical for evaluating an image's aesthetics. In this paper, we present an Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A -Lamp CNN) architecture for photo aesthetic assessment. This novel scheme is abl e to accept arbitrary sized images, and learn from both fined grained details an d holistic image layout simultaneously. To enable training on these hybrid input s, we extend the method by developing a dedicated double-subnet neural network s tructure, i.e. a Multi-Patch subnet and a Layout-Aware subnet. We further constr uct an aggregation layer to effectively combine the hybrid features from these t wo subnets. Extensive experiments on the large-scale aesthetics assessment bench mark (AVA) demonstrate significant performance improvement over the state-of-the -art in photo aesthetic assessment.
**********

Toroidal Constraints for Two-Point Localization Under High Outlier Ratios
Federico Camposeco, Torsten Sattler, Andrea Cohen, Andreas Geiger, Marc Pollefey s; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4545-4553
Localizing a query image against a 3D model at large scale is a hard problem, si nce 2D-3D matches become more and more ambiguous as the model size increases. Th is creates a need for pose estimation strategies that can handle very low inlier ratios. In this paper, we draw new insights on the geometric information availa ble from the 2D-3D matching process. As modern descriptors are not invariant aga inst large variations in viewpoint, we are able to find the rays in space used t o triangulate a given point that are closest to a query descriptor. It is well k nown that two correspondences constrain the camera to lie on the surface of a to rus. Adding the knowledge of direction of triangulation, we are able to approxim ate the position of the camera from two matches alone. We derive a geometric sol ver that can compute this position in under 1 microsecond. Using this solver, we propose a simple yet powerful outlier filter which scales quadratically in the number of matches. We validate the accuracy of our solver and demonstrate the us efulness of our method in real world settings.
**********

Hidden Layers in Perceptual Learning
Gad Cohen, Daphna Weinshall; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2017, pp. 4554-4562
Studies in visual perceptual learning investigate the way human performance impr oves with practice, in the context of relatively simple (and therefore more mana geable) visual tasks. Building on the powerful tools currently available for the training of Convolution Neural Networks (CNN), networks whose original architec ture was inspired by the visual system, we revisited some of the open computatio nal questions in perceptual learning. We first replicated two representative set s of perceptual learning experiments by training a shallow CNN to perform the re levant tasks. These networks qualitatively showed most of the characteristic beh avior observed in perceptual learning, including the hallmark phenomena of speci ficity and its various manifestations in the forms of transfer or partial transf er, and learning enabling. We next analyzed the dynamics of weight modifications in the networks, identifying patterns which appeared to be instrumental for the transfer (or generalization) of learned skills from one task to another in the simulated networks. These patterns may identify ways by which the domain of sear

ch in the parameter space during network re-training can be significantly reduced, thereby accomplishing knowledge transfer.
**********

InterpoNet, a Brain Inspired Neural Network for Optical Flow Dense Interpolation
Shay Zweig, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4563-4572
Sparse-to-dense interpolation for optical flow is a fundamental phase in the pipeline of most of the leading optical flow estimation algorithms. The current state-of-the-art method for interpolation, EpicFlow, is a local average method based on an edge aware geodesic distance. We propose a new data-driven sparse-to-dense interpolation algorithm based on a fully convolutional network. We draw inspiration from the filling-in process in the visual cortex and introduce lateral dependencies between neurons and multi-layer supervision into our learning process. We also show the importance of the image contour to the learning process. Our method is robust and outperforms EpicFlow on competitive optical flow benchmarks with several underlying matching algorithms. This leads to state-of-the-art performance on the Sintel and KITTI 2012 benchmarks.
**********

Learning Category-Specific 3D Shape Models From Weakly Labeled 2D Images
Dingwen Zhang, Junwei Han, Yang Yang, Dong Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4573-4581
Recently, researchers have made great processes to build category-specific 3D shape models from 2D images with manual annotations consisting of class labels, keypoints, and ground truth figure-ground segmentations. However, the annotation of figure-ground segmentations is still labor-intensive and time-consuming. To further alleviate the burden of providing such manual annotations, we make the earliest effort to learn category-specific 3D shape models by only using weakly labeled 2D images. By revealing the underlying relationship between the tasks of common object segmentation and category-specific 3D shape reconstruction, we propose a novel framework to jointly solve these two problems along a cluster-level learning curriculum. Comprehensive experiments on the challenging PASCAL VOC benchmark demonstrate that the category-specific 3D shape models trained using our weakly supervised learning framework could, to some extent, approach the performance of the state-of-the-art methods using expensive manual segmentation annotations. In addition, the experiments also demonstrate the effectiveness of using 3D shape models for helping common object segmentation.
**********

Zero-Shot Learning - the Good, the Bad and the Ugly
Yongqin Xian, Bernt Schiele, Zeynep Akata; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4582-4591
Due to the importance of zero-shot learning, the number of proposed approaches has increased steadily recently. We argue that it is time to take a step back and to analyze the status quo of the area. The purpose of this paper is three-fold. First, given the fact that there is no agreed upon zero-shot learning benchmark, we first define a new benchmark by unifying both the evaluation protocols and data splits. This is an important contribution as published results are often not comparable and sometimes even flawed due to, e.g. pre-training on zero-shot test classes. Second, we compare and analyze a significant number of the state-of-the-art methods in depth, both in the classic zero-shot setting but also in the more realistic generalized zero-shot setting. Finally, we discuss limitations of the current status of the area which can be taken as a basis for advancing it.
**********

Learning the Multilinear Structure of Visual Data
Mengjiao Wang, Yannis Panagakis, Patrick Snape, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4592-4600
Statistical decomposition methods are of paramount importance in discovering the modes of variations of visual data. Probably the most prominent linear decomposition method is the Principal Component Analysis (PCA), which discovers a single mode of variation in the data. However, in practice, visual data exhibit severa

l modes of variations. For instance, the appearance of faces varies in identity, expression, pose etc. To extract these modes of variations from visual data, several supervised methods, such as the TensorFaces, that rely on multilinear (tensor) decomposition (e.g., Higher Order SVD) have been developed. The main drawbacks of such methods is that they require both labels regarding the modes of variations and the same number of samples under all modes of variations (e.g., the same face under different expressions, poses etc.). Therefore, their applicability is limited to well-organised data, usually captured in well-controlled conditions. In this paper, we propose the first general multilinear method, to the best of our knowledge, that discovers the multilinear structure of visual data in unsupervised setting. That is, without the presence of labels. We demonstrate the applicability of the proposed method in two applications, namely Shape from Shading (SfS) and expression transfer.

**********

Linking Image and Text With 2-Way Nets

Aviv Eisenschtat, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4601-4611

Linking two data sources is a basic building block in numerous computer vision problems. Canonical Correlation Analysis (CCA) achieves this by utilizing a linear optimizer in order to maximize the correlation between the two views. Recent work makes use of non-linear models, including deep learning techniques, that optimize the CCA loss in some feature space. In this paper, we introduce a novel, bi-directional neural network architecture for the task of matching vectors from two data sources. Our approach employs two tied neural network channels that project the two views into a common, maximally correlated space using the Euclidean loss. We show a direct link between the correlation-based loss and Euclidean loss, enabling the use of Euclidean loss for correlation maximization. To overcome common Euclidean regression optimization problems, we modify well-known techniques to our problem, including batch normalization and dropout. We show state of the art results on a number of computer vision matching tasks including MNIST image matching and sentence-image matching on the Flickr8k, Flickr30k and COCO datasets.

**********

Unsupervised Semantic Scene Labeling for Streaming Data

Maggie Wigness, John G. Rogers III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4612-4621

We introduce an unsupervised semantic scene labeling approach that continuously learns and adapts semantic models discovered within a data stream. While closely related to unsupervised video segmentation, our algorithm is not designed to be an early video processing strategy that produces coherent over-segmentations, but instead, to directly learn higher-level semantic concepts. This is achieved with an ensemble-based approach, where each learner clusters data from a local window in the data stream. Overlapping local windows are processed and encoded in a graph structure to create a label mapping across windows and reconcile the labelings to reduce unsupervised learning noise. Additionally, we iteratively learn a merging threshold criteria from observed data similarities to automatically determine the number of learned labels without human provided parameters. Experiments show that our approach semantically labels video streams with a high degree of accuracy, and achieves a better balance of under and over-segmentation entropy than existing video segmentation algorithms given similar numbers of label outputs.

**********

Fast-At: Fast Automatic Thumbnail Generation Using Deep Neural Networks

Seyed A. Esmaeili, Bharat Singh, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4622-4630

Fast-AT is an automatic thumbnail generation system based on deep neural networks. It is a fully-convolutional deep neural network, which learns specific filters for thumbnails of different sizes and aspect ratios. During inference, the appropriate filter is selected depending on the dimensions of the target thumbnail. Unlike most previous work, Fast-AT does not utilize saliency but addresses the

problem directly. In addition, it eliminates the need to conduct region search over the saliency map. The model generalizes to thumbnails of different sizes including those with extreme aspect ratios and can generate thumbnails in real time. A data set of more than 70,000 thumbnail annotations was collected to train Fast-AT. We show competitive results in comparison to existing techniques.
**********

3D Point Cloud Registration for Localization Using a Deep Neural Network Auto-Encoder
Gil Elbaz, Tamar Avraham, Anath Fischer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4631-4640
We present an algorithm for registration between a large-scale point cloud and a close-proximity scanned point cloud, providing a localization solution that is fully independent of prior information about the initial positions of the two point cloud coordinate systems. The algorithm, denoted LORAX, selects super-points --local subsets of points--and describes the geometric structure of each with a low-dimensional descriptor. These descriptors are then used to infer potential matching regions for an efficient coarse registration process, followed by a fine-tuning stage. The set of super-points is selected by covering the point clouds with overlapping spheres, and then filtering out those of low-quality or nonsalient regions. The descriptors are computed using state-of-the-art unsupervised machine learning, utilizing the technology of deep neural network based auto-encoders.  Abstract This novel framework provides a strong alternative to the common practice of using manually designed key-point descriptors for coarse point cloud registration. Utilizing super-points instead of key-points allows the available geometrical data to be better exploited to find the correct transformation. Encoding local 3D geometric structures using a deep neural network auto-encoder instead of traditional descriptors continues the trend seen in other computer vision applications and indeed leads to superior results. The algorithm is tested on challenging point cloud registration datasets, and its advantages over previous approaches as well as its robustness to density changes, noise, and missing data are shown.
**********

Improved Stereo Matching With Constant Highway Networks and Reflective Confidence Learning
Amit Shaked, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4641-4650
We present an improved three-step pipeline for the stereo matching problem and introduce multiple novelties at each stage. We propose a new highway network architecture for computing the matching cost at each possible disparity, based on multilevel weighted residual shortcuts, trained with a hybrid loss that supports multilevel comparison of image patches. A novel post-processing step is then introduced, which employs a second deep convolutional neural network for pooling global information from multiple disparities. This network outputs both the image disparity map, which replaces the conventional "winner takes all" strategy, and a confidence in the prediction. The confidence score is achieved by training the network with a new technique that we call the reflective loss. Lastly, the learned confidence is employed in order to better detect outliers in the refinement step. The proposed pipeline achieves state of the art accuracy on the largest and most competitive stereo benchmarks, and the learned confidence is shown to outperform all existing alternatives.
**********

Superpixels and Polygons Using Simple Non-Iterative Clustering
Radhakrishna Achanta, Sabine Susstrunk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4651-4660
We present an improved version of the Simple Linear Iterative Clustering (SLIC) superpixel segmentation. Unlike SLIC, our algorithm is non-iterative, enforces connectivity from the start, requires lesser memory, and is faster. Relying on the superpixel boundaries obtained using our algorithm, we also present a polygonal partitioning algorithm. We demonstrate that our superpixels as well as the polygonal partitioning are superior to the respective state-of-the-art algorithms o

n quantitative benchmarks.
**********
POSEidon: Face-From-Depth for Driver Pose Estimation
Guido Borghi, Marco Venturelli, Roberto Vezzani, Rita Cucchiara; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4661-4670
Fast and accurate upper-body and head pose estimation is a key task for automatic monitoring of driver attention, a challenging context characterized by severe illumination changes, occlusions and extreme poses. In this work, we present a new deep learning framework for head localization and pose estimation on depth images. The core of the proposal is a regressive neural network, called POSEidon, which is composed of three independent convolutional nets followed by a fusion layer, specially conceived for understanding the pose by depth. In addition, to recover the intrinsic value of face appearance for understanding head position and orientation, we propose a new Face-from-Depth model for learning image faces from depth. Results in face reconstruction are qualitatively impressive. We test the proposed framework on two public datasets, namely Biwi Kinect Head Pose and ICT-3DHP, and on Pandora, a new challenging dataset mainly inspired by the automotive setup. Results show that our method overcomes all recent state-of-art works, running in real time at more than 30 frames per second.
**********
Optical Flow in Mostly Rigid Scenes
Jonas Wulff, Laura Sevilla-Lara, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4671-4680
The optical flow of natural scenes is a combination of the motion of the observer and the independent motion of objects. Existing algorithms typically focus on either recovering motion and structure under the assumption of a purely static world or optical flow for general unconstrained scenes. We combine these approaches in an optical flow algorithm that estimates an explicit segmentation of moving objects from appearance and physical constraints. In static regions we take advantage of strong constraints to jointly estimate the camera motion and the 3D structure of the scene over multiple frames. This allows us to also regularize the structure instead of the motion. Our formulation uses a Plane+Parallax framework, which works even under small baselines, and reduces the motion estimation to a one-dimensional search problem, resulting in more accurate estimation. In moving regions the flow is treated as unconstrained, and computed with an existing optical flow method. The resulting Mostly-Rigid Flow (MR-Flow) method achieves state-of-the-art results on both the MPI-Sintel and KITTI-2015 benchmarks.
**********
Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network
Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4681-4690
Despite the breakthroughs in accuracy and speed of single image super-resolution using faster and deeper convolutional neural networks, one central problem remains largely unsolved: how do we recover the finer texture details when we super-resolve at large upscaling factors? The behavior of optimization-based super-resolution methods is principally driven by the choice of the objective function. Recent work has largely focused on minimizing the mean squared reconstruction error. The resulting estimates have high peak signal-to-noise ratios, but they are often lacking high-frequency details and are perceptually unsatisfying in the sense that they fail to match the fidelity expected at the higher resolution. In this paper, we present SRGAN, a generative adversarial network (GAN) for image super-resolution (SR). To our knowledge, it is the first framework capable of inferring photo-realistic natural images for 4x upscaling factors. To achieve this, we propose a perceptual loss function which consists of an adversarial loss and a content loss. The adversarial loss pushes our solution to the natural image manifold using a discriminator network that is trained to differentiate between th

e super-resolved images and original photo-realistic images. In addition, we use a content loss motivated by perceptual similarity instead of similarity in pixel space. Our deep residual network is able to recover photo-realistic textures from heavily downsampled images on public benchmarks. An extensive mean-opinion-score (MOS) test shows hugely significant gains in perceptual quality using SRGAN. The MOS scores obtained with SRGAN are closer to those of the original high-resolution images than to those obtained with any state-of-the-art method.
**********

ROAM: A Rich Object Appearance Model With Application to Rotoscoping
Ondrej Miksik, Juan-Manuel Perez-Rua, Philip H. S. Torr, Patrick Perez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4691-4699
Rotoscoping, the detailed delineation of scene elements through a video shot, is a painstaking task of tremendous importance in professional post-production pipelines. While pixel-wise segmentation techniques can help for this task, professional rotoscoping tools rely on parametric curves that offer the artists a much better interactive control on the definition, editing and manipulation of the segments of interest. Sticking to this prevalent rotoscoping paradigm, we propose a novel framework to capture and track the visual aspect of an arbitrary object in a scene, given a first closed outline of this object. This model combines a collection of local foreground/background appearance models spread along the outline, a global appearance model of the enclosed object and a set of distinctive foreground landmarks. The structure of this rich appearance model allows simple initialization, efficient iterative optimization with exact minimization at each step, and on-line adaptation in videos. We demonstrate qualitatively and quantitatively the merit of this framework through comparisons with tools based on either dynamic segmentation with a closed curve or pixel-wise binary labelling.
**********

Densely Connected Convolutional Networks
Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708
Recent work has shown that convolutional networks can be substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. In this paper, we embrace this observation and introduce the Dense Convolutional Network (DenseNet), which connects each layer to every other layer in a feed-forward fashion. Whereas traditional convolutional networks with L layers have L connections--one between each layer and its subsequent layer--our network has L(L+1)/2 direct connections. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. We evaluate our proposed architecture on four highly competitive object recognition benchmark tasks (CIFAR- 10, CIFAR-100, SVHN, and ImageNet). DenseNets obtain significant improvements over the state-of-the-art on most of them, whilst requiring less memory and computation to achieve high performance. Code and pre-trained models are available at https://github.com/liuzhuang13/DenseNet.
**********

Multi-Level Attention Networks for Visual Question Answering
Dongfei Yu, Jianlong Fu, Tao Mei, Yong Rui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4709-4717
Inspired by the recent success of text-based question answering, visual question answering (VQA) is proposed to automatically answer natural language questions with the reference to a given image. Compared with text-based QA, VQA is more challenging because the reasoning process on visual domain needs both effective semantic embedding and fine-grained visual understanding. Existing approaches predominantly infer answers from the abstract low-level visual features, while neglecting the modeling of high-level image semantics and the rich spatial context of

regions. To solve the challenges, we propose a multi-level attention network for visual question answering that can simultaneously reduce the semantic gap by semantic attention and benefit fine-grained spatial inference by visual attention. First, we generate semantic concepts from high-level semantics in convolutional neural networks (CNN) and select those question-related concepts as semantic attention. Second, we encode region-based middle-level outputs from CNN into spatially-embedded representation by a bidirectional recurrent neural network, and further pinpoint the answer-related regions by multiple layer perceptron as visual attention. Third, we jointly optimize semantic attention, visual attention and question embedding by a softmax classifier to infer the final answer. Extensive experiments show the proposed approach outperforms the-state-of-arts on two challenging VQA datasets.

**********

Spatial-Semantic Image Search by Visual Feature Synthesis

Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4718-4727

The performance of image retrieval has been improved tremendously in recent years through the use of deep feature representations. Most existing methods, however, aim to retrieve images that are visually similar or semantically relevant to the query, irrespective of spatial configuration. In this paper, we develop a spatial-semantic image search technology that enables users to search for images with both semantic and spatial constraints by manipulating concept text-boxes on a 2D query canvas. We train a convolutional neural network to synthesize appropriate visual features that captures the spatial-semantic constraints from the user canvas query. We directly optimize the retrieval performance of the visual features when training our deep neural network. These visual features then are used to retrieve images that are both spatially and semantically relevant to the user query. The experiments on large-scale datasets such as MS-COCO and Visual Genome show that our method outperforms other baseline and state-of-the-art methods in spatial-semantic image search.

**********

Temporal Residual Networks for Dynamic Scene Recognition

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4728-4737

This paper combines three contributions to establish a new state-of-the-art in dynamic scene recognition. First, we present a novel ConvNet architecture based on temporal residual units that is fully convolutional in spacetime. Our model augments spatial ResNets with convolutions across time to hierarchically add temporal residuals as the depth of the network increases. Second, existing approaches to video-based recognition are categorized and a baseline of seven previously top performing algorithms is selected for comparative evaluation on dynamic scenes. Third, we introduce a new and challenging video database of dynamic scenes that more than doubles the size of those previously available. This dataset is explicitly split into two subsets of equal size that contain videos with and without camera motion to allow for systematic study of how this variable interacts with the defining dynamics of the scene per se. Our evaluations verify the particular strengths and weaknesses of the baseline algorithms with respect to various scene classes and camera motion parameters. Finally, our temporal ResNet boosts recognition performance and establishes a new state-of-the-art on dynamic scene recognition, as well as on the complementary task of action recognition.

**********

Radiometric Calibration for Internet Photo Collections

Zhipeng Mo, Boxin Shi, Sai-Kit Yeung, Yasuyuki Matsushita; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4738-4746

Radiometrically calibrating the images from Internet photo collections brings photometric analysis from lab data to big image data in the wild, but conventional calibration methods cannot be directly applied to such image data. This paper presents a method to jointly perform radiometric calibration for a set of images

in an Internet photo collection. By incorporating the consistency of scene reflectance for corresponding pixels in multiple images, the proposed method estimates radiometric response functions of all the images using a rank minimization framework. Our calibration aligns all response functions in an image set up to the same exponential ambiguity in a robust manner. Quantitative results using both synthetic and real data show the effectiveness of the proposed method.
**********

See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-Identification
Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, Tieniu Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4747-4756
Surveillance cameras have been widely used in different scenes. Accordingly, a demanding need is to recognize a person under different cameras, which is called person re-identification. This topic has gained increasing interests in computer vision recently. However, less attention has been paid to video-based approaches, compared with image-based ones. Two steps are usually involved in previous approaches, namely feature learning and metric learning. But most of the existing approaches only focus on either feature learning or metric learning. Meanwhile, many of them do not take full use of the temporal and spatial information. In this paper, we concentrate on video-based person re-identification and build an end-to-end deep neural network architecture to jointly learn features and metrics. The proposed method can automatically pick out the most discriminative frames in a given video by a temporal attention model. Moreover, it integrates the surrounding information at each location by a spatial recurrent model when measuring the similarity with another pedestrian video. That is, our method handles spatial and temporal information simultaneously in a unified manner. The carefully designed experiments on three public datasets show the effectiveness of each component of the proposed deep network, performing better in comparison with the state-of-the-art methods.
**********

Procedural Generation of Videos to Train Deep Action Recognition Networks
Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, Antonio Manuel Lopez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4757-4767
Deep learning for human action recognition in videos is making significant progress, but is slowed down by its dependency on expensive manual labeling of large video collections. In this work, we investigate the generation of synthetic training data for action recognition, as it has recently shown promising results for a variety of other computer vision tasks. We propose an interpretable parametric generative model of human action videos that relies on procedural generation and other computer graphics techniques of modern game engines. We generate a diverse, realistic, and physically plausible dataset of human action videos, called PHAV for "Procedural Human Action Videos". It contains a total of 39,982 videos, with more than 1,000 examples for each action of 35 categories. Our approach is not limited to existing motion capture sequences, and we procedurally define 14 synthetic actions. We introduce a deep multi-task representation learning architecture to mix synthetic and real videos, even if the action categories differ. Our experiments on the UCF101 and HMDB51 benchmarks suggest that combining our large set of synthetic videos with small real-world datasets can boost recognition performance, significantly outperforming fine-tuning state-of-the-art unsupervised generative models of videos.
**********

Spatiotemporal Multiplier Networks for Video Action Recognition
Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4768-4777
This paper presents a general ConvNet architecture for video action recognition based on multiplicative interactions of spacetime features. Our model combines the appearance and motion pathways of a two-stream architecture by motion gating and is trained end-to-end. We theoretically motivate multiplicative gating funct

ions for residual networks and empirically study their effect on classification accuracy. To capture long-term dependencies we inject identity mapping kernels for learning temporal relationships. Our architecture is fully convolutional in spacetime and able to evaluate a video in a single forward pass. Empirical inves tigation reveals that our model produces state-of-the-art results on two standar d action recognition datasets.

**********

Real-Time Video Super-Resolution With Spatio-Temporal Networks and Motion Compen sation

Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, Wenzhe Shi; Proceedings of the IEEE Conference on Computer Vision a nd Pattern Recognition (CVPR), 2017, pp. 4778-4787

Convolutional neural networks have enabled accurate image super-resolution in re al-time. However, recent attempts to benefit from temporal correlations in video super-resolution have been limited to naive or inefficient architectures. In th is paper, we introduce spatio-temporal sub-pixel convolution networks that effec tively exploit temporal redundancies and improve reconstruction accuracy while m aintaining real-time speed. Specifically, we discuss the use of early fusion, sl ow fusion and 3D convolutions for the joint processing of multiple consecutive v ideo frames. We also propose a novel joint motion compensation and video super-r esolution algorithm that is orders of magnitude more efficient than competing me thods, relying on a fast multi-resolution spatial transformer module that is end -to-end trainable. These contributions provide both higher accuracy and temporal ly more consistent videos, which we confirm qualitatively and quantitatively. Re lative to single-frame models, spatio-temporal networks can either reduce the co mputational cost by 30% whilst maintaining the same quality or provide a 0.2dB g ain for a similar computational cost. Results on publicly available datasets dem onstrate that the proposed algorithms surpass current state-of-the-art performan ce in both accuracy and efficiency.

**********

Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Bas ed Approach

Aidean Sharghi, Jacob S. Laurel, Boqing Gong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4788-4797

Recent years have witnessed a resurgence of interest in video summarization. How ever, one of the main obstacles to the research on video summarization is the us er subjectivity --- users have various preferences over the summaries. The subje ctiveness causes at least two problems. First, no single video summarizer fits a ll users unless it interacts with and adapts to the individual users. Second, it is very challenging to evaluate the performance of a video summarizer. To tac kle the first problem, we explore the recently proposed query-focused video summ arization which introduces user preferences in the form of text queries about t he video into the summarization process. We propose a memory network parameteriz ed sequential determinantal point process in order to attend the user query onto different video frames and shots. To address the second challenge, we contend t hat a good evaluation metric for video summarization should focus on the semanti c information that humans can perceive rather than the visual features or tempor al overlaps. To this end, we collect dense per-video-shot concept annotations, c ompile a new dataset, and suggest an efficient evaluation method defined upon t he concept annotations. We conduct extensive experiments contrasting our video summarizer to existing ones and present detailed analyses about the dataset and the new evaluation method.

**********

A New Rank Constraint on Multi-View Fundamental Matrices, and Its Application to Camera Location Recovery

Soumyadip Sengupta, Tal Amir, Meirav Galun, Tom Goldstein, David W. Jacobs, Amit Singer, Ronen Basri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4798-4806

Accurate estimation of camera matrices is an important step in structure from mo tion algorithms. In this paper we introduce a novel rank constraint on collectio

ns of fundamental matrices in multi-view settings. We show that in general, with the selection of proper scale factors, a matrix formed by stacking fundamental matrices between pairs of images has rank 6. Moreover, this matrix forms the symmetric part of a rank 3 matrix whose factors relate directly to the corresponding camera matrices. We use this new characterization to produce better estimations of fundamental matrices by optimizing an L1-cost function using Iterative Re-weighted Least Squares and Alternate Direction Method of Multiplier. We further show that this procedure can improve the recovery of camera locations, particularly in multi-view settings in which fewer images are available.

**********

Attentional Correlation Filter Network for Adaptive Visual Tracking

Jongwon Choi, Hyung Jin Chang, Sangdoo Yun, Tobias Fischer, Yiannis Demiris, Jin Young Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4807-4816

We propose a new tracking framework with an attentional mechanism that chooses a subset of the associated correlation filters for increased robustness and computational efficiency. The subset of filters is adaptively selected by a deep attentional network according to the dynamic properties of the tracking target. Our contributions are manifold, and are summarised as follows: (i) Introducing the Attentional Correlation Filter Network which allows adaptive tracking of dynamic targets. (ii) Utilising an attentional network which shifts the attention to the best candidate modules, as well as predicting the estimated accuracy of currently inactive modules. (iii) Enlarging the variety of correlation filters which cover target drift, blurriness, occlusion, scale changes, and flexible aspect ratio. (iv) Validating the robustness and efficiency of the attentional mechanism for visual tracking through a number of experiments. Our method achieves similar performance to non real-time trackers, and state-of-the-art performance amongst real-time trackers.

**********

Deep Mixture of Linear Inverse Regressions Applied to Head-Pose Estimation

Stephane Lathuiliere, Remi Juge, Pablo Mesejo, Rafael Munoz-Salinas, Radu Horaud; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4817-4825

Convolutional Neural Networks (ConvNets) have become the state-of-the-art for many classification and regression problems in computer vision. When it comes to regression, approaches such as measuring the Euclidean distance of target and predictions are often employed as output layer. In this paper, we propose the coupling of a Gaussian mixture of linear inverse regressions with a ConvNet, and we describe the methodological foundations and the associated algorithm to jointly train the deep network and the regression function. We test our model on the head-pose estimation problem. In this particular problem, we show that inverse regression outperforms regression models currently used by state-of-the-art computer vision methods. Our method does not require the incorporation of additional data, as it is often proposed in the literature, thus it is able to work well on relatively small training datasets. Finally, it outperforms state-of-the-art methods in head-pose estimation using a widely used head-pose dataset. To the best of our knowledge, we are the first to incorporate inverse regression into deep learning for computer vision applications.

**********

Human Shape From Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks

Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, Markus Gross; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4826-4836

In this work, we present a novel method for capturing human body shape from a single scaled silhouette. We combine deep correlated features capturing different 2D views, and embedding spaces based on 3D cues in a novel convolutional neural network (CNN) based architecture. We first train a CNN to find a richer body shape representation space from pose invariant 3D human shape descriptors. Then, we learn a mapping from silhouettes to this representation space, with the help of

a novel architecture that exploits correlation of multi-view data during traini
ng time, to improve prediction at test time. We extensively validate our results
 on synthetic and real data, demonstrating significant improvements in accuracy
as compared to the state-of-the-art, and providing a practical system for detail
ed human body measurements from a single image.
**********

STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling
Yang He, Wei-Chen Chiu, Margret Keuper, Mario Fritz; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4837-4846
We propose a novel superpixel-based multi-view convolutional neural network for
semantic image segmentation. The proposed network produces a high quality segmen
tation of a single image by leveraging information from additional views of the
same scene. Particularly in indoor videos such as captured by robotic platforms
or handheld and bodyworn RGBD cameras, nearby video frames provide diverse viewp
oints and additional context of objects and scenes. To leverage such information
, we first compute region correspondences by optical flow and image boundary-bas
ed superpixels. Given these region correspondences, we propose a novel spatio-te
mporal pooling layer to aggregate information over space and time. We evaluate o
ur approach on the NYU-Depth-V2 and the SUN3D datasets and compare it to various
 state-of-the-art single-view and multi-view approaches. Besides a general impro
vement over the state-of-the-art, we also show the benefits of making use of unl
abeled frames during training for multi-view as well as single-view prediction.
**********

On the Two-View Geometry of Unsynchronized Cameras
Cenek Albl, Zuzana Kukelova, Andrew Fitzgibbon, Jan Heller, Matej Smid, Tomas Pa
jdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2017, pp. 4847-4856
We present new methods of simultaneously estimating camera geometry and time shi
ft from video sequences from multiple unsynchronized cameras.  Algorithms for si
multaneous computation of a fundamental matrix or a homography with unknown time
 shift between images are developed. Our methods use minimal correspondence sets
  (eight for fundamental matrix and four and a half for homography) and therefor
e are suitable for robust estimation using RANSAC. Furthermore, we present an it
erative algorithm that extends the applicability on sequences which are signific
antly unsynchronized, finding the correct time shift up to several seconds. We e
valuated the methods on synthetic and wide range of real world datasets and the
results show a broad applicability to the problem of camera synchronization.
**********

Using Locally Corresponding CAD Models for Dense 3D Reconstructions From a Singl
e Image
Chen Kong, Chen-Hsuan Lin, Simon Lucey; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2017, pp. 4857-4865
We investigate the problem of estimating the dense 3D shape of an object, given
a set of 2D landmarks and silhouette in a single image. An obvious prior to empl
oy in such a problem is a dictionary of dense CAD models. Employing a sufficient
ly large enough dictionary of CAD models, however, is in general computationally
 infeasible. A common strategy in dictionary learning to encourage generalizatio
n is to allow for linear combinations of dictionary elements. This too, however,
 is problematic as most CAD models cannot be readily placed in global dense corr
espondence. In this paper, we propose a two-step strategy. First, we employ orth
ogonal matching pursuit to rapidly choose the "closest" single CAD model in our
dictionary to the projected image. Second, we employ a novel graph embedding bas
ed on local dense correspondence to allow for sparse linear combinations of CAD
models. We validate our framework experimentally in both synthetic and real worl
d scenario and demonstrate the superiority of our approach to both 3D mesh recon
struction and volumetric representation.
**********

BigHand2.2M Benchmark: Hand Pose Dataset and State of the Art Analysis
Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, Tae-Kyun Kim; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.

4866-4874
In this paper we introduce a large-scale hand pose dataset, collected using a novel capture method. Existing datasets are either generated synthetically or captured using depth sensors: synthetic datasets exhibit a certain level of appearance difference from real depth images, and real datasets are limited in quantity and coverage, mainly due to the difficulty to annotate them. We propose a tracking system with six 6D magnetic sensors and inverse kinematics to automatically obtain 21-joints hand pose annotations of depth maps captured with minimal restriction on the range of motion. The capture protocol aims to fully cover the natural hand pose space. As shown in embedding plots, the new dataset exhibits a significantly wider and denser range of hand poses compared to existing benchmarks. Current state-of-the-art methods are evaluated on the dataset, and we demonstrate significant improvements in cross-benchmark performance. We also show significant improvements in egocentric hand pose estimation with a CNN trained on the new dataset.
**********
A Matrix Splitting Method for Composite Function Minimization
Ganzhao Yuan, Wei-Shi Zheng, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4875-4884
Composite function minimization captures a wide spectrum of applications in both computer vision and machine learning. It includes bound constrained optimization and cardinality regularized optimization as special cases. This paper proposes and analyzes a new Matrix Splitting Method (MSM) for minimizing composite functions. It can be viewed as a generalization of the classical Gauss-Seidel method and the Successive Over-Relaxation method for solving linear systems in the literature. Incorporating a new Gaussian elimination procedure, the matrix splitting method achieves state-of-the-art performance. For convex problems, we establish the global convergence, convergence rate, and iteration complexity of MSM, while for non-convex problems, we prove its global convergence. Finally, we validate the performance of our matrix splitting method on two particular applications: nonnegative matrix factorization and cardinality regularized sparse coding. Extensive experiments show that our method outperforms existing composite function minimization techniques in term of both efficiency and efficacy.
**********
Simultaneous Geometric and Radiometric Calibration of a Projector-Camera Pair
Marjan Shahpaski, Luis Ricardo Sapaico, Gaspard Chevassus, Sabine Susstrunk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4885-4893
We present a novel method that allows for simultaneous geometric and radiometric calibration of a projector-camera pair. It is simple and does not require specialized hardware. We prewarp and align a specially designed projection pattern onto a printed pattern of different colorimetric properties. After capturing the patterns in several orientations, we perform geometric calibration by estimating the corner locations of the two patterns in different color channels. We perform radiometric calibration of the projector by using the information contained inside the projected squares. We show that our method performs on par with current approaches that all require separate geometric and radiometric calibration, while being more efficient and user friendly.
**********
On the Global Geometry of Sphere-Constrained Sparse Blind Deconvolution
Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, John Wright; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4894-4902
Blind deconvolution is the problem of recovering a convolutional kernel and an activation signal from their convolution. This problem is ill-posed without further constraints or priors. This paper studies the situation where the nonzero entries in the activation signal are sparsely and randomly populated.. We normalize the convolution kernel to have unit Frobenius norm and cast the sparse blind deconvolution problem as a nonconvex optimization problem over the sphere. With this spherical constraint, every spurious local minimum turns out to be close to s

ome signed shift truncation of the ground truth, under certain hypotheses. This benign property motivates an effective two stage algorithm that recovers the ground truth from the partial information offered by a suboptimal local minimum. This geometry-inspired algorithm recovers the ground truth for certain microscopy problems, also exhibits promising performance in the more challenging image deblurring problem. Our insights into the global geometry and the two stage algorithm extend to the convolutional dictionary learning problem, where a superposition of multiple convolution signals is observed.
**********

Towards Accurate Multi-Person Pose Estimation in the Wild
George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, Kevin Murphy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4903-4911
We propose a method for multi-person detection and 2-D pose estimation that achieves state-of-art results on the challenging COCO keypoints task. It is a simple, yet powerful, top-down approach consisting of two stages.  In the first stage, we predict the location and scale of boxes which are likely to contain people; for this we use the Faster RCNN detector. In the second stage, we estimate the keypoints of the person potentially contained in each proposed bounding box. For each keypoint type we predict dense heatmaps and offsets using a fully convolutional ResNet. To combine these outputs we introduce a novel aggregation procedure to obtain highly localized keypoint predictions. We also use a novel form of keypoint-based Non-Maximum-Suppression (NMS), instead of the cruder box-level NMS, and a novel form of keypoint-based confidence score estimation, instead of box-level scoring.  Trained on COCO data alone, our final system achieves average precision of 0.649 on the COCO test-dev set and the 0.643 test-standard sets, outperforming the winner of the 2016 COCO keypoints challenge and other recent state-of-art. Further, by using additional in-house labeled data we obtain an even higher average precision of 0.685 on the test-dev set and 0.673 on the test-standard set, more than 5% absolute improvement compared to the previous best performing method on the same dataset.
**********

A Clever Elimination Strategy for Efficient Minimal Solvers
Zuzana Kukelova, Joe Kileel, Bernd Sturmfels, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4912-4921
We present a new insight into the systematic generation of minimal solvers in computer vision, which leads to smaller and faster solvers. Many minimal problem formulations are coupled sets of linear and polynomial equations where image measurements enter the linear equations only. We show that it is useful to solve such systems by first eliminating all the unknowns that do not appear in the linear equations and then extending solutions to the rest of unknowns. This can be generalized to fully non-linear systems by linearization via lifting. We demonstrate that this approach leads to more efficient solvers in three problems of partially calibrated relative camera pose computation with unknown focal length and/or radial distortion. Our approach also generates new interesting constraints on the fundamental matrices of partially calibrated cameras, which were not known before.
**********

Adaptive and Move Making Auxiliary Cuts for Binary Pairwise Energies
Lena Gorelick, Yuri Boykov, Olga Veksler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4922-4930
Many computer vision problems require optimization of binary non-submodular energies. In this context, iterative submodularization techniques based on trust region (LSA-TR) and auxiliary functions (LSA-AUX) have been recently proposed [??]. They achieve state-of-the-art-results on a number of computer vision applications.  In this paper we extend the LSA-AUX framework in two directions. First, unlike LSA-AUX which selects auxiliary functions based solely on the current solution, we propose to incorporate several additional criteria. This results in tighter bounds for configurations that are more likely or closer to the current solut

ion. Second, we propose move-making extensions of LSA-AUX which achieve tighter bounds by restricting the search space. Finally, we evaluate our methods on several applications. We show that for each application at least one of our extensions significantly outperforms the original LSA-AUX. Moreover, the best extension of LSA-AUX is comparable to or better than LSA-TR on five out of six applications, achieving state-of-the-arts results on four out of six applications.
**********

What Is the Space of Attenuation Coefficients in Underwater Computer Vision?
Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, David Iluz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4931-4940
Underwater image reconstruction methods require the knowledge of wideband attenuation coefficients per color channel. Current estimation methods for these coefficients require specialized hardware or multiple images, and none of them leverage the multitude of existing ocean optical measurements as priors. Here, we aim to constrain the set of physically-feasible wideband attenuation coefficients in the ocean by utilizing water attenuation measured worldwide by oceanographers. We calculate the space of valid wideband effective attenuation coefficients in the 3D RGB domain and find that a bound manifold in 3-space sufficiently represents the variation from the clearest to murkiest waters. We validate our model using in situ experiments in two different optical water bodies, the Red Sea and the Mediterranean. Moreover, we show that contradictory to the common image formation model, the coefficients depend on the imaging range and object reflectance, and quantify the errors resulting from ignoring these dependencies.
**********

Consensus Maximization With Linear Matrix Inequality Constraints
Pablo Speciale, Danda Pani Paudel, Martin R. Oswald, Till Kroeger, Luc Van Gool, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4941-4949
Consensus maximization has proven to be a useful tool for robust estimation. While randomized methods like RANSAC are fast, they do not guarantee global optimality and fail to manage large amounts of outliers. On the other hand, global methods are commonly slow because they do not exploit the structure of the problem at hand. In this paper, we show that the solution space can be reduced by introducing Linear Matrix Inequality (LMI) constraints. This leads to significant speed ups of the optimization time even for large amounts of outliers, while maintaining global optimality. We study several cases in which the objective variables have a special structure, such as rotation, scaled-rotation, and essential matrices, which are posed as LMI constraints. This is very useful in several standard computer vision problems, such as estimating Similarity Transformations, Absolute Poses, and Relative Poses, for which we obtain compelling results on both synthetic and real datasets. With up to 90 percent outlier rate, where RANSAC often fails, our constrained approach is consistently faster than the non-constrained one - while finding the same global solution.
**********

Optical Flow Requires Multiple Strategies (but Only One Network)
Tal Schuster, Lior Wolf, David Gadot; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4950-4959
We show that the matching problem that underlies optical flow requires multiple strategies, depending on the amount of image motion and other factors. We then study the implications of this observation on training a deep neural network for representing image patches in the context of descriptor based optical flow. We propose a metric learning method, which selects suitable negative samples based on the nature of the true match. This type of training produces a network that displays multiple strategies depending on the input and leads to state of the art results on the KITTI 2012 and KITTI 2015 optical flow benchmarks.
**********

Convex Global 3D Registration With Lagrangian Duality
Jesus Briales, Javier Gonzalez-Jimenez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4960-4969

The registration of 3D models by a Euclidean transformation is a fundamental tas k at the core of many application in computer vision. This problem is non-convex due to the presence of rotational constraints, making traditional local optimiz ation methods prone to getting stuck in local minima. This paper addresses findi ng the globally optimal transformation in various 3D registration problems by a unified formulation that integrates common geometric registration modalities (na mely point-to-point, point-to-line and point-to-plane). This formulation renders the optimization problem independent of both the number and nature of the corre spondences. The main novelty of our proposal is the introduction of a strengthe ned Lagrangian dual relaxation for this problem, which surpasses previous simila r approaches [32] in effectiveness. In fact, even though with no theoretical gua rantees, exhaustive empirical evaluation in both synthetic and real experiments always resulted on a tight relaxation that allowed to recover a guaranteed globa lly optimal solution by exploiting duality theory. Thus, our approach allows fo r effectively solving the 3D registration with global optimality guarantees whil e running at a fraction of the time for the state-of-the-art alternative [34], b ased on a more computationally intensive Branch and Bound method. [32] C. Olsso n and A. Eriksson, "Solving Quadratically Constrained Geometrical Problems using Lagrangian Duality," in Proc. 19th Int. Conf. Pattern Recognition (ICPR2008), p p. 1-5. [34] C. Olsson, F. Kahl, and M. Oskarsson. "Branch-and-Bound Methods for Euclidean Registration Problems," in IEEE Trans. Pattern Anal. Mach. Intell., 3 1(5):783-794, 2009.

**********

S3Pool: Pooling With Stochastic Spatial Sampling

Shuangfei Zhai, Hui Wu, Abhishek Kumar, Yu Cheng, Yongxi Lu, Zhongfei Zhang, Rog erio Feris; Proceedings of the IEEE Conference on Computer Vision and Pattern Re cognition (CVPR), 2017, pp. 4970-4978

Feature pooling layers (e.g., max pooling) in convolutional neural networks (CNN s) serve the dual purpose of providing increasingly abstract representations as well as yielding computational savings in subsequent convolutional layers. We vi ew the pooling operation in CNNs as a two step procedure: first, a pooling windo w (e.g., 2x2) slides over the feature map with stride one which leaves the spati al resolution intact, and second, downsampling is performed by selecting one pi xel from each non-overlapping pooling window in an often uniform and determinist ic (e.g., top-left) manner. Our starting point in this work is the observation t hat this regularly spaced downsampling arising from non-overlapping windows, alt hough intuitive from a signal processing perspective (which has the goal of sign al reconstruction), is not necessarily optimal for learning (where the goal is t o generalize). We study this aspect and propose a novel pool ing strategy with stochastic spatial sampling (S3Pool), where the regular downsa mpling is replaced by a more general stochastic version. We observe that th is general stochasticity acts as a strong regularizer, and can also be seen as d oing implicit data augmentation by introducing distortions in the feature maps. We further introduce a mechanism to control the amount of distortion to suit di fferent datasets and architectures. To demonstrate the effectiveness of the prop osed approach, we perform extensive experiments on several popular image classif ication benchmarks, observing excellent improvements over baseline models.

**********

Generating Descriptions With Grounded and Co-Referenced People

Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, Bernt Schiele; Proceed ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2 017, pp. 4979-4989

Learning how to generate descriptions of images or videos received major interes t both in the Computer Vision and Natural Language Processing communities. While a few works have proposed to learn a grounding during the generation process in an unsupervised way (via an attention mechanism), it remains unclear how good t he quality of the grounding is and whether it benefits the description quality. In this work we propose a movie description model which learns to generate descr iption and jointly ground (localize) the mentioned characters as well as do visu al co-reference resolution between pairs of consecutive sentences/clips. We also

propose to use weak localization supervision through character mentions provided in movie descriptions to learn the character grounding. At training time, we first learn how to localize characters by relating their visual appearance to mentions in the descriptions via a semi-supervised approach. We then provide this (noisy) supervision into our description model which greatly improves its performance. Our proposed description model improves over prior work w.r.t. generated description quality and additionally provides grounding and local co-reference resolution. We evaluate it on the MPII Movie Description dataset using automatic and human evaluation measures and using our newly collected grounding and co-reference data for characters.
**********

Deep Photo Style Transfer
Fujun Luan, Sylvain Paris, Eli Shechtman, Kavita Bala; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4990-4998
This paper introduces a deep-learning approach to photographic style transfer that handles a large variety of image content while faithfully transferring the reference style. Our approach builds upon the recent work on painterly transfer that separates style from the content of an image by considering different layers of a neural network. However, as is, this approach is not suitable for photorealistic style transfer. Even when both the input and reference images are photographs, the output still exhibits distortions reminiscent of a painting. Our contribution is to constrain the transformation from the input to the output to be locally affine in colorspace, and to express this constraint as a custom fully differentiable energy term. We show that this approach successfully suppresses distortion and yields satisfying photorealistic style transfers in a broad variety of scenarios, including transfer of the time of day, weather, season, and artistic edits.
**********

Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension
Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, Hannaneh Hajishirzi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4999-5007
We introduce the task of Multi-Modal Machine Comprehension (M3C), which aims at answering multimodal questions given a context of text, diagrams and images. We present the Textbook Question Answering (TQA) dataset that includes 1,076 lessons and 26,260 multi-modal questions, taken from middle school science curricula. Our analysis shows that a significant portion of questions require complex parsing of the text and the diagrams and reasoning, indicating that our dataset is more complex compared to previous machine comprehension and visual question answering datasets. We extend state-of-the-art methods for textual machine comprehension and visual question answering to the TQA dataset. Our experiments show that these models do not perform well on TQA. The presented dataset opens new challenges for research in question answering and reasoning across multiple modalities.
**********

InstanceCut: From Edges to Instances With MultiCut
Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5008-5017
This work addresses the task of instance-aware semantic segmentation. Our key motivation is to design a simple method with a new modelling-paradigm, which therefore has a different trade-off between advantages and disadvantages compared to known approaches. Our approach, we term InstanceCut, represents the problem by two output modalities: (i) an instance-agnostic semantic segmentation and (ii) all instance-boundaries. The former is computed from a standard convolutional neural network for semantic segmentation, and the latter is derived from a new instance-aware edge detection model. To reason globally about the optimal partitioning of an image into instances, we combine these two modalities into a novel MultiCut formulation. We evaluate our approach on the challenging CityScapes dataset. Despite the conceptual simplicity of our approach, we achieve the best result a

mong all published methods, and perform particularly well for rare object classes.
**********

Deep Hashing Network for Unsupervised Domain Adaptation
Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, Sethuraman Panchanathan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5018-5027
In recent years, deep neural networks have emerged as a dominant machine learning tool for a wide variety of application domains. However, training a deep neural network requires a large amount of labeled data, which is an expensive process in terms of time, labor and human expertise. Domain adaptation or transfer learning algorithms address this challenge by leveraging labeled data in a different, but related source domain, to develop a model for the target domain. Further, the explosive growth of digital data has posed a fundamental challenge concerning its storage and retrieval. Due to its storage and retrieval efficiency, recent years have witnessed a wide application of hashing in a variety of computer vision applications. In this paper, we first introduce a new dataset, Office-Home, to evaluate domain adaptation algorithms. The dataset contains images of a variety of everyday objects from multiple domains. We then propose a novel deep learning framework that can exploit labeled source data and unlabeled target data to learn informative hash codes, to accurately classify unseen target data. To the best of our knowledge, this is the first research effort to exploit the feature learning capabilities of deep neural networks to learn representative hash codes to address the domain adaptation problem. Our extensive empirical studies on multiple transfer tasks corroborate the usefulness of the framework in learning efficient hash codes which outperform existing competitive baselines for unsupervised domain adaptation.
**********

Harmonic Networks: Deep Translation and Rotation Equivariance
Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, Gabriel J. Brostow; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5028-5037
Translating or rotating an input image should not affect the results of many computer vision tasks. Convolutional neural networks (CNNs) are already translation equivariant: input image translations produce proportionate feature map translations. This is not the case for rotations. Global rotation equivariance is typically sought through data augmentation, but patch-wise equivariance is more difficult. We present Harmonic Networks or H-Nets, a CNN exhibiting equivariance to patch-wise translation and 360-rotation. We achieve this by replacing regular CNN filters with circular harmonics, returning a maximal response and orientation for every receptive field patch.  H-Nets use a rich, parameter-efficient and fixed computational complexity representation, and we show that deep feature maps within the network encode complicated rotational invariants. We demonstrate that our layers are general enough to be used in conjunction with the latest architectures and techniques, such as deep supervision and batch normalization. We also achieve state-of-the-art classification on rotated-MNIST, and competitive results on other benchmark challenges.
**********

DeMoN: Depth and Motion Network for Learning Monocular Stereo
Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, Thomas Brox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5038-5047
In this paper we formulate structure from motion as a learning problem. We train a convolutional network end-to-end to compute depth and camera motion from successive, unconstrained image pairs. The architecture is composed of multiple stacked encoder-decoder networks, the core part being an iterative network that is able to improve its own predictions. The network estimates not only depth and motion, but additionally surface normals, optical flow between the images and confidence of the matching. A crucial component of the approach is a training loss based on spatial relative differences. Compared to traditional two-frame structure

from motion methods, results are more accurate and more robust. In contrast to the popular depth-from-single-image networks, DeMoN learns the concept of matching and, thus, better generalizes to structures not seen during training.
**********

A Wide-Field-Of-View Monocentric Light Field Camera
Donald G. Dansereau, Glenn Schuster, Joseph Ford, Gordon Wetzstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5048-5057
Light field (LF) capture and processing are important in an expanding range of computer vision applications, offering rich textural and depth information and simplification of conventionally complex tasks. Although LF cameras are commercially available, no existing device offers wide field-of-view (FOV) imaging. This is due in part to the limitations of fisheye lenses, for which a fundamentally constrained entrance pupil diameter severely limits depth sensitivity. In this work we describe a novel, compact optical design that couples a monocentric lens with multiple sensors using microlens arrays, allowing LF capture with an unprecedented FOV. Leveraging capabilities of the LF representation, we propose a novel method for efficiently coupling the spherical lens and planar sensors, replacing expensive and bulky fiber bundles. We construct a single-sensor LF camera prototype, rotating the sensor relative to a fixed main lens to emulate a wide-FOV multi-sensor scenario. Finally, we describe a processing toolchain, including a convenient spherical LF parameterization, and demonstrate depth estimation and post-capture refocus for indoor and outdoor panoramas with 15 x 15 x 1600 x 200 pixels (72 MPix) and a 138-degree FOV.
**********

Teaching Compositionality to CNNs
Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D. Scott Phoenix, Dileep George; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5058-5067
Convolutional neural networks (CNNs) have shown great success in computer vision, approaching human-level performance when trained for specific tasks via application- specific loss functions. In this paper, we propose a method for augmenting and training CNNs so that their learned features are compositional. It encourages networks to form representations that disentangle objects from their surroundings and from each other, thereby promoting better generalization. Our method is agnostic to the specific de- tails of the underlying CNN to which it is applied and can in principle be used with any CNN. As we show in our experiments, the learned representations lead to feature activations that are more localized and improve performance over non-compositional baselines in object recognition tasks.
**********

Learning Barycentric Representations of 3D Shapes for Sketch-Based 3D Shape Retrieval
Jin Xie, Guoxian Dai, Fan Zhu, Yi Fang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5068-5076
Retrieving 3D shapes with sketches is a challenging problem since 2D sketches and 3D shapes are from two heterogeneous domains, which results in large discrepancy between them. In this paper, we propose to learn barycenters of 2D projections of 3D shapes for sketch-based 3D shape retrieval. Specifically, we first use two deep convolutional neural networks (CNNs) to extract deep features of sketches and 2D projections of 3D shapes. For 3D shapes, we then compute the Wasserstein barycenters of deep features of multiple projections to form a barycentric representation. Finally, by constructing a metric network, a discriminative loss is formulated on the Wasserstein barycenters of 3D shapes and sketches in the deep feature space to learn discriminative and compact 3D shape and sketch features for retrieval. The proposed method is evaluated on the SHREC'13 and SHREC'14 sketch track benchmark datasets. Compared to the state-of-the-art methods, our proposed method can significantly improve the retrieval performance.
**********

Stacked Generative Adversarial Networks

Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5077-5086

In this paper, we propose a novel generative model named Stacked Generative Adversarial Networks (SGAN), which is trained to invert the hierarchical representations of a bottom-up discriminative network. Our model consists of a top-down stack of GANs, each learned to generate lower-level representations conditioned on higher-level representations. A representation discriminator is introduced at each feature hierarchy to encourage the representation manifold of the generator to align with that of the bottom-up discriminative network, leveraging the powerful discriminative representations to guide the generative model. In addition, we introduce a conditional loss that encourages the use of conditional information from the layer above, and a novel entropy loss that maximizes a variational lower bound on the conditional entropy of generator outputs. We first train each stack independently, and then train the whole model end-to-end. Unlike the original GAN that uses a single noise vector to represent all the variations, our SGAN decomposes variations into multiple levels and gradually resolves uncertainties in the top-down generative process. Based on visual inspection, Inception scores and visual Turing test, we demonstrate that SGAN is able to generate images of much higher quality than GANs without stacking.

**********

Image Splicing Detection via Camera Response Function Analysis

Can Chen, Scott McCloskey, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5087-5096

Recent advances on image manipulation techniques have made image forgery detection increasingly more challenging. An important component in such tools is to fake motion and/or defocus blurs through boundary splicing and copy-move operators, to emulate wide aperture and slow shutter effects. In this paper, we present a new technique based on the analysis of the camera response functions (CRF) for efficient and robust splicing and copy-move forgery detection and localization. We first analyze how non-linear CRFs affect edges in terms of the intensity-gradient bivariable histograms. We show distinguishable shape differences on real vs. forged blurs near edges after a splicing operation. Based on our analysis, we introduce a deep-learning framework to detect and localize forged edges. In particular, we show the problem can be transformed to a handwriting recognition problem an resolved by using a convolutional neural network. We generate a large data set of forged images produced by splicing followed by retouching and comprehensive experiments show our proposed method outperforms the state-of-the-art techniques in accuracy and robustness.

**********

Illuminant-Camera Communication to Observe Moving Objects Under Strong External Light by Spread Spectrum Modulation

Ryusuke Sagawa, Yutaka Satoh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5097-5105

Many algorithms of computer vision use light sources to illuminate objects to actively create situation appropriate to extract their characteristics. For example, the shape and reflectance are measured by a projector-camera system, and some human-machine or VR systems use projectors and displays for interaction. As existing active lighting systems usually assume no severe external lights to observe projected lights clearly, it is one of the limitations of active illumination. In this paper, we propose a method of energy-efficient active illumination in an environment with severe external lights. The proposed method extracts the light signals of illuminants by removing external light using spread spectrum modulation. Because an image sequence is needed to observe modulated signals, the proposed method extends signal processing to realize signal detection projected onto moving objects by combining spread spectrum modulation and spatio-temporal filtering. In the experiments, we apply the proposed method to a structured-light system under sunlight, to photometric stereo with external lights, and to insensible image embedding.

**********

Building a Regular Decision Boundary With Deep Networks
Edouard Oyallon; Proceedings of the IEEE Conference on Computer Vision and Patte rn Recognition (CVPR), 2017, pp. 5106-5114
In this work, we build a generic architecture of Convolutional Neural Networks t o discover empirical properties of neural networks. Our first contribution is to introduce a state-of-the-art framework that depends upon few hyper parameters a nd to study the network when we vary them. It has no max pooling, no biases, onl y 13 layers, is purely convolutional and yields up to 95.4% and 79.6% accuracy respectively on CIFAR10 and CIFAR100. We show that the nonlinearity of a deep ne twork does not need to be continuous, non expansive or point-wise, to achieve go od performance. We show that increasing the width of our network permits being c ompetitive with very deep networks. Our second contribution is an analysis of th e contraction and separation properties of this network. Indeed, a 1-nearest nei ghbor classifier applied on deep features progressively improves with depth, whi ch indicates that the representation is progressively more regular. Besides, we defined and analyzed local support vectors that separate classes locally. All o ur experiments are reproducible and code will be available online, based on Tens orFlow.
**********

Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs
Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, M ichael M. Bronstein; Proceedings of the IEEE Conference on Computer Vision and P attern Recognition (CVPR), 2017, pp. 5115-5124
Deep learning has achieved a remarkable performance breakthrough in several fiel ds, most notably in speech recognition, natural language processing, and compute r vision. In particular, convolutional neural network (CNN) architectures curren tly produce state-of-the-art performance on a variety of image analysis tasks su ch as object detection and recognition. Most of deep learning research has so fa r focused on dealing with 1D, 2D, or 3D Euclidean-structured data such as acoust ic signals, images, or videos. Recently, there has been an increasing interest i n geometric deep learning, attempting to generalize deep learning methods to non -Euclidean structured data such as graphs and manifolds, with a variety of appli cations from the domains of network analysis, computational social science, or c omputer graphics. In this paper, we propose a unified framework allowing to gene ralize CNN architectures to non-Euclidean domains (graphs and manifolds) and lea rn local, stationary, and compositional task-specific features. We show that var ious non-Euclidean CNN methods previously proposed in the literature can be cons idered as particular instances of our framework. We test the proposed method on standard tasks from the realms of image-, graph- and 3D shape analysis and show that it consistently outperforms previous approaches.
**********

Identifying First-Person Camera Wearers in Third-Person Videos
Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J. Crandall, Michael S. Ryoo; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2017, pp. 5125-5133
We consider scenarios in which we wish to perform joint scene understand ing, object tracking, activity recognition, and other tasks in scenarios in wh ich multiple people are wearing body-worn cameras while a third-person static ca mera also captures the scene. To do this, we need to establish person-level cor respondences across first- and third-person videos, which is challenging because the camera wearer is not visible from his/her own egocentric video, preventing the use of direct feature matching. In this paper, we propose a new semi-Siamese Convolutional Neural Network architecture to address this novel challenge. We formulate the problem as learning a joint embedding space for first- and third-person videos that considers both spatial- and motion-doma in cues. A new triplet loss function is designed to minimize the distance betwe en correct first- and third-person matches while maximizing the distance betw een incorrect ones. This end-to-end approach performs significantly better than several baselines, in part by learning the first- and third-person feature s optimized for matching jointly with the distance measure itself.

**********
IM2CAD
Hamid Izadinia, Qi Shan, Steven M. Seitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5134-5143
Given a single photo of a room and a large database of furniture CAD models, our goal is to reconstruct a scene that is as similar as possible to the scene depicted in the photograph, and composed of objects drawn from the database. We present a completely automatic system to address this IM2CAD problem that produces high quality results on challenging imagery from interior home design and remodeling websites. Our approach iteratively optimizes the placement and scale of objects in the room to best match scene renderings to the input photo, using image comparison metrics trained via deep convolutional neural nets. By operating jointly on the full scene at once, we account for inter-object occlusions. We also show the applicability of our method in standard scene understanding benchmarks where we obtain significant improvement.
**********
Photorealistic Facial Texture Inference Using Deep Neural Networks
Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, Hao Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5144-5153
We present a data-driven inference method that can synthesize a photorealistic texture map of a complete 3D face model given a partial 2D view of a person in the wild.  After an initial estimation of shape and low-frequency albedo, we compute a high-frequency partial texture map, without the shading component, of the visible face area.  To extract the fine appearance details from this incomplete input, we introduce a multi-scale detail analysis technique based on mid-layer feature correlations extracted from a deep convolutional neural network.  We demonstrate that fitting a convex combination of feature correlations from a high-resolution face database can yield a semantically plausible facial detail description of the entire face. A complete and photorealistic texture map can then be synthesized by iteratively optimizing for the reconstructed feature correlations. Using these high-resolution textures and a commercial rendering framework, we can produce high-fidelity 3D renderings that are visually comparable to those obtained with state-of-the-art multi-view face capture systems. We demonstrate successful face reconstructions from a wide range of low resolution input images, including those of historical figures. In addition to extensive evaluations, we validate the realism of our results using a crowdsourced user study.
**********
Learning to Learn From Noisy Web Videos
Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5154-5162
Understanding the simultaneously very diverse and intricately fine-grained set of possible human actions is a critical open problem in computer vision. Manually labeling training videos is feasible for some action classes but doesn't scale to the full long-tailed distribution of actions. A promising way to address this is to leverage noisy data from web queries to learn new actions, using semi-supervised or "webly-supervised" approaches. However, these methods typically do not learn domain-specific knowledge, or rely on iterative hand-tuned data labeling policies. In this work, we instead propose a reinforcement learning-based formulation for selecting the right examples for training a classifier from noisy web search results. Our method uses Q-learning to learn a data labeling policy on a small labeled training dataset, and then uses this to automatically label noisy web data for new visual concepts. Experiments on the challenging Sports-1M action recognition benchmark as well as on additional fine-grained and newly emerging action classes demonstrate that our method is able to learn good labeling policies for noisy data and use this to learn accurate visual concept classifiers.
**********
Regressing Robust and Discriminative 3D Morphable Models With a Very Deep Neural Network

Anh Tuan Tran, Tal Hassner, Iacopo Masi, Gerard Medioni; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5163-5172

The 3D shapes of faces are well known to be discriminative. Yet despite this, they are rarely used for face recognition and always under controlled viewing conditions. We claim that this is a symptom of a serious but often overlooked problem with existing methods for single view 3D face reconstruction: when applied "in the wild", their 3D estimates are either unstable and change for different photos of the same subject or they are over-regularized and generic. In response, we describe a robust method for regressing discriminative 3D morphable face models (3DMM). We use a convolutional neural network (CNN) to regress 3DMM shape and texture parameters directly from an input photo. We overcome the shortage of training data required for this purpose by offering a method for generating huge numbers of labeled examples. The 3D estimates produced by our CNN surpass state of the art accuracy on the MICC data set. Coupled with a 3D-3D face matching pipeline, we show the first competitive face recognition results on the LFW, YTF and IJB-A benchmarks using 3D face shapes as representations, rather than the opaque deep feature vectors used by other modern systems.

**********

HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, Krystian Mikolajczyk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5173-5182

In this paper, we propose a novel benchmark for evaluating local image descriptors. We demonstrate that the existing datasets and evaluation protocols do not specify unambiguously all aspects of evaluation, leading to ambiguities and inconsistencies in results reported in the literature. Furthermore, these datasets are nearly saturated due to the recent improvements in local descriptors obtained by learning them from large annotated datasets. Therefore, we introduce a new large dataset suitable for training and testing modern descriptors, together with strictly defined evaluation protocols in several tasks such as matching, retrieval and classification. This allows for more realistic, and thus more reliable comparisons in different application scenarios. We evaluate the performance of several state-of-the-art descriptors and analyse their properties. We show that a simple normalisation of traditional hand-crafted descriptors can boost their performance to the level of deep learning based descriptors within a realistic benchmarks evaluation.

**********

Using Ranking-CNN for Age Estimation

Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, Mike Rao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5183-5192

Human age is considered an important biometric trait for human identification or search. Recent research shows that the aging features deeply learned from large-scale data lead to significant performance improvement on facial image-based age estimation. However, age-related ordinal information is totally ignored in these approaches. In this paper, we propose a novel Convolutional Neural Network (CNN)-based framework, ranking-CNN, for age estimation. Ranking-CNN contains a series of basic CNNs, each of which is trained with ordinal age labels. Then, their binary outputs are aggregated for the final age prediction. We theoretically obtain a much tighter error bound for ranking-based age estimation. Moreover, we rigorously prove that ranking-CNN is more likely to get smaller estimation errors when compared with multi-class classification approaches. Through extensive experiments, we show that statistically, ranking-CNN significantly outperforms other state-of-the-art age estimation models on benchmark datasets.

**********

DeepNav: Learning to Navigate Large Cities

Samarth Brahmbhatt, James Hays; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5193-5202

We present DeepNav, a Convolutional Neural Network (CNN) based algorithm for navigating large cities using locally visible street-view images. The DeepNav agent learns to reach its destination quickly by making the correct navigation decisions at intersections. We collect a large-scale dataset of street-view images organized in a graph where nodes are connected by roads. This dataset contains 10 city graphs and a total of more than 1 million street-view images. We propose 3 supervised learning approaches for the navigation task, and show how A* search in the city graph can be used to generate labels for the images. Our annotation process is fully automated using publicly available mapping services, and requires no human input. We evaluate the proposed DeepNav models on 4 held-out cities for navigating to 5 different types of destinations and show that our algorithms outperform previous work that uses hand-crafted features and Support Vector Regression (SVR).

**********

The World of Fast Moving Objects

Denys Rozumnyi, Jan Kotera, Filip Sroubek, Lukas Novotny, Jiri Matas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5203-5211

The notion of a Fast Moving Object (FMO), i.e. an object that moves over a distance exceeding its size within the exposure time, is introduced. FMOs may, and typically do, rotate with high angular speed. FMOs are very common in sports videos, but are not rare elsewhere. In a single frame, such objects are often barely visible and appear as semitransparent streaks. A method for the detection and tracking of FMOs is proposed. The method consists of three distinct algorithms, which form an efficient localization pipeline that operates successfully in a broad range of conditions. We show that it is possible to recover the appearance of the object and its axis of rotation, despite its blurred appearance. The proposed method is evaluated on a new annotated dataset. The results show that existing trackers are inadequate for the problem of FMO localization and a new approach is required. Two applications of localization, temporal superresolution and highlighting, are presented.

**********

Sports Field Localization via Deep Structured Models

Namdar Homayounfar, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5212-5220

In this work, we propose a novel way of efficiently localizing a sports field from a single broadcast image of the game. Related work in this area relies on manually annotating a few key frames and extending the localization to similar images, or installing fixed specialized cameras in the stadium from which the layout of the field can be obtained. In contrast, we formulate this problem as a branch and bound inference in a Markov random field where an energy function is defined in terms of  semantic cues such as the field surface, lines and circles obtained from a deep semantic segmentation network. Moreover, our approach is fully automatic and depends only on a single image from the broadcast video of the game. We demonstrate the effectiveness of our method by applying it to soccer and hockey.

**********

Deep Watershed Transform for Instance Segmentation

Min Bai, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5221-5229

Most contemporary approaches to instance segmentation use complex pipelines involving conditional random fields, recurrent neural networks, object proposals, or template matching schemes. In this paper, we present a simple yet powerful end-to-end convolutional neural network to tackle this task. Our approach combines intuitions from the classical watershed transform and modern deep learning to produce an energy map of the image where object instances are unambiguously represented as energy basins. We then perform a cut at a single energy level to directly yield connected components corresponding to object instances. Our model achieves more than double the performance over the state-of-the-art on the challenging Cityscapes Instance Level Segmentation task.

**********
Annotating Object Instances With a Polygon-RNN
Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5230-5238
In this paper, we propose an approach for semi-automatic annotation of object instances. While most current methods treat object segmentation as a pixel-labeling problem, we here cast it as a polygon prediction task, mimicking how most current datasets have been annotated. In particular, our approach takes as input an image crop and produces a vertex of the polygon, one at a time, allowing the human annotator to interfere at any time and correct the point. Our model easily integrates any correction, producing as accurate segmentations as desired by the annotator. We show that our annotation method speeds up the annotation process by a factor of 4.7 across all classes, while achieving 78.4% agreement in IoU with original ground-truth, matching the typical agreement between human annotators. For cars, our speed-up factor is even higher, at 7.3 for agreement of 82.2%. We further show generalization capabilities of our approach on unseen datasets.
**********
Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer
Xin Wang, Geoffrey Oxholm, Da Zhang, Yuan-Fang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5239-5247
Transferring artistic styles onto everyday photographs has become an extremely popular task in both academia and industry. Recently, offline training has replaced online iterative optimization, enabling nearly real-time stylization. When those stylization networks are applied directly to high-resolution images, however, the style of localized regions often appears less similar to the desired artistic style. This is because the transfer process fails to capture small, intricate textures and maintain correct texture scales of the artworks. Here we propose a multimodal convolutional neural network that takes into consideration faithful representations of both color and luminance channels, and performs stylization hierarchically with multiple losses of increasing scales. Compared to state-of-the-art networks, our network can also perform style transfer in nearly real-time by performing much more sophisticated training offline. By properly handling style and texture cues at multiple scales using several modalities, we can transfer not just large-scale, obvious style cues but also subtle, exquisite ones. That is, our scheme can generate results that are visually pleasing and more similar to multiple desired artistic styles with color and texture cues at multiple scales.
**********
Detect, Replace, Refine: Deep Structured Prediction for Pixel Wise Labeling
Spyros Gidaris, Nikos Komodakis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5248-5257
Pixel wise image labeling is an interesting and challenging problem with great significance in the computer vision community. In order for a dense labeling algorithm to be able to achieve accurate and precise results, it has to consider the dependencies that exist in the joint space of both the input and the output variables. An implicit approach for modeling those dependencies is by training a deep neural network that, given as input an initial estimate of the output labels and the input image, it will be able to predict a new refined estimate for the labels. In this context, our work is concerned with what is the optimal architecture for performing the label improvement task. We argue that the prior approaches of either directly predicting new label estimates or predicting residual corrections w.r.t. the initial labels with feed-forward deep network architectures are sub-optimal. Instead, we propose a generic architecture that decomposes the label improvement task to three steps: 1) detecting the initial label estimates that are incorrect, 2) replacing the incorrect labels with new ones, and finally 3) refining the renewed labels by predicting residual corrections w.r.t. them. Furthermore, we explore and compare various other alternative architectures that consist of the aforementioned Detection, Replace, and Refine components. We exten

sively evaluate the examined architectures in the challenging task of dense disp
arity estimation (stereo matching) and we report both quantitative and qualitati
ve results on three different datasets. Finally, our dense disparity estimation
network that implements the proposed generic architecture, achieves state-of-the
-art results in the KITTI 2015 test surpassing prior approaches by a significant
 margin. We plan to release the Torch code that implements the paper in: https:/
/github.com/gidariss/DRR_struct_pred/ .
**********

Grassmannian Manifold Optimization Assisted Sparse Spectral Clustering
Qiong Wang, Junbin Gao, Hong Li; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2017, pp. 5258-5266
Spectral Clustering is one of pioneered clustering methods in machine learning a
nd pattern recognition field. It relies on the spectral decomposition criterion
to learn a low-dimensonal embedding of data for a basic clustering algorithm suc
h as the k-means. The recent sparse Spectral clustering (SSC) introduces the spa
rsity for the similarity in low-dimensional space by enforcing a sparsity-induce
d penalty, resulting a non-convex optimization, and the solution is calculated t
hrough a relaxed convex problem via the standard ADMM (Alternative Direction Met
hod of Multipliers), rather than inferring latent representation from eigen-stru
cture. This paper provides a direct solution as solving a new Grassmann optimiza
tion problem. By this way calculating latent embedding becomes part of optmizati
on on manifolds and the recently developed manifold optimization methods can be
applied.  It turns out the learned new features are not only very informative f
or clustering, but also more intuitive and effective in visualization after dime
nsionality reduction. We conduct empirical studies on simulated datasets and sev
eral real-world benchmark datasets to validate the proposed methods. Experimenta
l results exhibit the effectiveness of this new manifold-based clustering and di
mensionality reduction method.
**********

Robust Joint and Individual Variance Explained
Christos Sagonas, Yannis Panagakis, Alina Leidinger, Stefanos Zafeiriou; Proceed
ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2
017, pp. 5267-5276
Discovering the common (joint) and individual subspaces is crucial for analysis
of multiple data sets, including multi-view and multi-modal data. Several statis
tical machine learning methods have been developed for discovering the common fe
atures across multiple data sets. The most well studied family of the methods is
 that of Canonical Correlation Analysis (CCA) and its variants. Even though the
CCA is a powerful tool, it has several drawbacks that render its application cha
llenging for computer vision applications. That is, it discovers only common fea
tures and not individual ones, and it is sensitive to gross errors present in vi
sual data.  Recently, efforts have been made in order to develop methods that d
iscover individual and common components. Nevertheless, these methods are mainly
 applicable in two sets of data. In this paper, we investigate the use of a rece
ntly proposed statistical  method, the so-called Joint and Individual Variance E
xplained (JIVE) method, for the recovery of joint and individual components in a
n arbitrary number of data sets. Since, the JIVE is not robust to gross errors,
we propose alternatives, which are both robust to non-Gaussian noise of large ma
gnitude, as well as able to automatically find the rank of the individual compon
ents. We demonstrate the effectiveness of the proposed approach to two computer
vision applications, namely facial expression synthesis and face age progression
 in-the-wild.
**********

AnchorNet: A Weakly Supervised Network to Learn Geometry-Sensitive Features for
Semantic Matching
David Novotny, Diane Larlus, Andrea Vedaldi; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5277-5286
Despite significant progress of deep learning in recent years, state-of-the-art
semantic matching methods still rely on legacy features such as SIFT or HoG. We
argue that the strong invariance properties that are key to the success of recen

t deep architectures on the classification task make them unfit for dense corres pondence tasks, unless a large amount of supervision is used. In this work, we p ropose a deep network, termed AnchorNet, that produces image representations tha t are well-suited for semantic matching. It relies on a set of filters whose res ponse is geometrically consistent across different object instances, even in the presence of strong intra-class, scale, or viewpoint variations. Trained only wi th weak image-level labels, the final representation successfully captures infor mation about the object structure and improves results of state-of-the-art seman tic matching methods such as the Deformable Spatial Pyramid or the Proposal Flow methods. We show positive results on the cross-instance matching task where dif ferent instances of the same object category are matched as well as on a new cro ss-category semantic matching task aligning pairs of instances each from a diffe rent object class.

**********

Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Ne ural Networks

Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin , Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and P attern Recognition (CVPR), 2017, pp. 5287-5295

Indoor scene understanding is central to applications such as robot navigation a nd human companion assistance. Over the last years, data-driven deep neural netw orks have outperformed many traditional approaches thanks to their representatio n learning capabilities. One of the bottlenecks in training for better represent ations is the amount of available per-pixel ground truth data that is required f or core scene understanding tasks such as semantic segmentation, normal predicti on, and object boundary detection. To address this problem, a number of works pr oposed using synthetic data. However, a systematic study of how such synthetic d ata is generated is missing. In this work, we introduce a large-scale synthetic dataset with 500K physically-based rendered images from 45K realistic 3D indoor scenes. We study the effects of rendering methods and scene lighting on trainin g for three computer vision tasks: surface normal prediction, semantic segmentat ion, and object boundary detection. This study provides insights into the best p ractices for training with synthetic data (more realistic rendering is worth it) and shows that pretraining with our new synthetic dataset can improve results b eyond the current state of the art on all three tasks.

**********

YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Objec t Detection in Video

Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, Vincent Vanhoucke; Pr oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP R), 2017, pp. 5296-5305

We introduce a new large-scale data set of video URLs with densely-sampled objec t bounding box annotations called YouTube-BoundingBoxes (YT-BB). The data set co nsists of approximately 380,000 video segments about 19s long, automatically sel ected to feature objects in natural settings without editing or post-processing, with a recording quality often akin to that of a hand-held cell phone camera. T he objects represent a subset of the COCO label set. All video segments were hum an-annotated with high-precision classification labels and bounding boxes at 1 f rame per second. The use of a cascade of increasingly precise human annotations ensures a label accuracy above 95% for every class and tight bounding boxes. Fin ally, we train and evaluate well-known deep network architectures and report bas eline figures for per-frame classification and localization. We also demonstrate how the temporal contiguity of video can potentially be used to improve such in ferences. The data set can be found at https://research.google.com/youtube-bb. W e hope the availability of such large curated corpus will spur new advances in v ideo object detection and tracking.

**********

Full Resolution Image Compression With Recurrent Neural Networks

George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Jo el Shor, Michele Covell; Proceedings of the IEEE Conference on Computer Vision a

nd Pattern Recognition (CVPR), 2017, pp. 5306-5314
This paper presents a set of full-resolution lossy image compression methods based on neural networks. Each of the architectures we describe can provide variable compression rates during deployment without requiring retraining of the network: each network need only be trained once. All of our architectures consist of a recurrent neural network (RNN)-based encoder and decoder, a binarizer, and a neural network for entropy coding. We compare RNN types (LSTM, associative LSTM) and introduce a new hybrid of GRU and ResNet. We also study "one-shot" versus additive reconstruction architectures and introduce a new scaled-additive framework. We compare to previous work, showing improvements of 4.3%-8.8% AUC (area under the rate-distortion curve), depending on the perceptual metric used. As far as we know, this is the first neural network architecture that is able to outperform JPEG at image compression across most bitrates on the rate-distortion curve on the Kodak dataset images, with and without the aid of entropy coding.
**********

Learning to Extract Semantic Structure From Documents Using Multimodal Fully Convolutional Neural Networks
Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, C. Lee Giles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5315-5324
We present an end-to-end, multimodal, fully convolutional network for extracting semantic structures from document images. We consider document semantic structure extraction as a pixel-wise segmentation task, and propose a unified model that classifies pixels based not only on their visual appearance, as in the traditional page segmentation task, but also on the content of underlying text. Moreover, we propose an efficient synthetic document generation process that we use to generate pretraining data for our network. Once the network is trained on a large set of synthetic documents, we fine-tune the network on unlabeled real documents using a semi-supervised approach. We systematically study the optimum network architecture and show that both our multimodal approach and the synthetic data pretraining significantly boost the performance.
**********

Hardware-Efficient Guided Image Filtering for Multi-Label Problem
Longquan Dai, Mengke Yuan, Zechao Li, Xiaopeng Zhang, Jinhui Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5325-5333
The Guided Filter (GF) is well-known for its linear complexity. However, when filtering an image with an n-channel guidance, GF needs to invert an n xn matrix for each pixel. To the best of our knowledge existing matrix inverse algorithms are inefficient on current hardwares. This shortcoming limits applications of multichannel guidance in computation intensive system such as multi-label system.
We need a new GF-like filter that can perform fast multichannel image guided filtering. Since the optimal linear complexity of GF cannot be minimized further, the only way thus is to bring all potentialities of current parallel computing hardwares into full play. In this paper we propose a hardware-efficient Guided Filter (HGF), which solves the efficiency problem of multichannel guided image filtering and yields competent results when applying it to multi-label problems with synthesized polynomial multichannel guidance. Specifically, in order to boost the filtering performance, HGF takes a new matrix inverse algorithm which only involves two hardware-efficient operations: element-wise arithmetic calculations and box filtering. In order to break the linear model restriction, HGF synthesizes a polynomial multichannel guidance to introduce nonlinearity. Benefiting from our polynomial guidance and hardware-efficient matrix inverse algorithm, HGF not only is more sensitive to the underlying structure of guidance but also achieves the fastest computing speed. Due to these merits, HGF obtains state-of-the-art results in terms of accuracy and efficiency in the computation intensive multi-label systems.
**********

Fully-Adaptive Feature Sharing in Multi-Task Networks With Applications in Person Attribute Classification

Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, Rogerio Feris; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5334-5343

Multi-task learning aims to improve generalization performance of multiple prediction tasks by appropriately sharing relevant information across them. In the context of deep neural networks, this idea is often realized by hand-designed network architectures with layers that are shared across tasks and branches that encode task-specific features. However, the space of possible multi-task deep architectures is combinatorially large and often the final architecture is arrived at by manual exploration of this space, which can be both error-prone and tedious. We propose an automatic approach for designing compact multi-task deep learning architectures. Our approach starts with a thin multi-layer network and dynamically widens it in a greedy manner during training. By doing so iteratively, it creates a tree-like deep architecture, on which similar tasks reside in the same branch until at the top layers. Evaluation on person attributes classification tasks involving facial and clothing attributes suggests that the models produced by the proposed method are fast, compact and can closely match or exceed the state-of-the-art accuracy from strong baselines by much more expensive models.

**********

Hyperspectral Image Super-Resolution via Non-Local Sparse Tensor Factorization

Renwei Dian, Leyuan Fang, Shutao Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5344-5353

Hyperspectral image(HSI)super-resolution, which fuses a low-resolution (LR) HSI with a high-resolution (HR) multispectral image (MSI), has recently attracted much attention. Most of the current HSI super-resolution approaches are based on matrix factorization, which unfolds the three-dimensional HSI as a matrix before processing. In general, the matrix data representation obtained after the matrix unfolding operation makes it hard to fully exploit the inherent HSI spatial-spectral structures. In this paper, a novel HSI super-resolution method based on non-local sparse tensor factorization (called as the NLSTF) is proposed. The sparse tensor factorization can directly decompose each cube of the HSI as a sparse core tensor and dictionaries of three modes, which reformulates the HSI super-resolution problem as the estimation of sparse core tensor and dictionaries for each cube. To further exploit the non-local spatial self-similarities of the HSI, similar cubes are grouped together, and they are assumed to share the same dictionaries. The dictionaries are learned from the LR-HSI and HR-MSI for each group, and corresponding sparse core tensors are estimated by spare coding on the learned dictionaries for each cube. Experimental results demonstrate the superiority of the proposed NLSTF approach over several state-of-the-art HSI super-resolution approaches.

**********

Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation

Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5354-5362

This paper addresses the problem of depth estimation from a single still image. Inspired by recent works on multi-scale convolutional neural networks (CNN), we propose a deep model which fuses complementary information derived from multiple CNN side outputs. Different from previous methods, the integration is obtained by means of continuous Conditional Random Fields (CRFs). In particular, we propose two different variations, one based on a cascade of multiple CRFs, the other on a unified graphical model. By designing a novel CNN implementation of mean-field updates for continuous CRFs, we show that both proposed models can be regarded as sequential deep networks and that training can be performed end-to-end. Through extensive experimental evaluation we demonstrate the effectiveness of the proposed approach and establish new state of the art results on publicly available datasets.

**********

Learning Cross-Modal Deep Representations for Robust Pedestrian Detection

Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5363-5371

This paper presents a novel method for detecting pedestrians under adverse illumination conditions. Our approach relies on a novel cross-modality learning framework and it is based on two main phases. First, given a multimodal dataset, a deep convolutional network is employed to learn a non-linear mapping, modeling the relations between RGB and thermal data. Then, the learned feature representations are transferred to a second deep network, which receives as input an RGB image and outputs the detection results. In this way, features which are both discriminative and robust to bad illumination conditions are learned. Importantly, at test time, only the second pipeline is considered and no thermal data are required. Our extensive evaluation demonstrates that the proposed approach outperforms the state-of-the-art on the challenging KAIST multispectral pedestrian dataset and it is competitive with previous methods on the popular Caltech dataset.

**********

Noisy Softmax: Improving the Generalization Ability of DCNN via Postponing the Early Softmax Saturation

Binghui Chen, Weihong Deng, Junping Du; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5372-5381

Over the past few years, softmax and SGD have become a commonly used component and the default training strategy in CNN frameworks, respectively. However, when optimizing CNNs with SGD, the saturation behavior behind softmax always gives us an illusion of training well and then is omitted. In this paper, we first emphasize that the early saturation behavior of softmax will impede the exploration of SGD, which sometimes is a reason for model converging at a bad local-minima, then propose Noisy Softmax to mitigating this early saturation issue by injecting annealed noise in softmax during each iteration. This operation based on noise injection aims at postponing the early saturation and further bringing continuous gradients propagation so as to significantly encourage SGD solver to be more exploratory and help to find a better local-minima. This paper empirically verifies the superiority of the early softmax desaturation, and our method indeed improves the generalization ability of CNN model by regularization. We experimentally find that this early desaturation helps optimization in many tasks, yielding state-of-the-art or competitive results on several popular benchmark datasets.

**********

Deep Metric Learning via Facility Location

Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, Kevin Murphy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5382-5390

Learning image similarity metrics in an end-to-end fashion with deep networks has demonstrated excellent results on tasks such as clustering and retrieval. However, current methods, all focus on a very local view of the data. In this paper, we propose a new metric learning scheme, based on structured prediction, that is aware of the global structure of the embedding space, and which is designed to optimize a clustering quality metric (NMI). We show state of the art performance on standard datasets, such as CUB200-2011, Cars196, and Stanford online products on NMI and R@K evaluation metrics.

**********

Event-Based Visual Inertial Odometry

Alex Zihao Zhu, Nikolay Atanasov, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5391-5399

Event-based cameras provide a new visual sensing model by detecting changes in image intensity asynchronously across all pixels on the camera. By providing these events at extremely high rates (up to 1MHz), they allow for sensing in both high speed and high dynamic range situations where traditional cameras may fail. In this paper, we present the first algorithm to fuse a purely event-based tracking algorithm with an inertial measurement unit, to provide accurate metric tracking of a camera's full 6dof pose. Our algorithm is asynchronous, and provides measurement updates at a rate proportional to the camera velocity. The algorithm s

elects features in the image plane, and tracks spatiotemporal windows around the se features within the event stream. An Extended Kalman Filter with a structurel ess measurement model then fuses the feature tracks with the output of the IMU. The camera poses from the filter are then used to initialize the next step of th e tracker and reject failed tracks. We show that our method successfully tracks camera motion on the Event-Camera Dataset in a number of challenging situations.
**********

Scribbler: Controlling Deep Image Synthesis With Sketch and Color
Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, James Hays; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5400-5409
Recently, there have been several promising methods to generate realistic imager y from deep convolutional networks. These methods sidestep the traditional compu ter graphics rendering pipeline and instead generate imagery at the pixel level by learning from large collections of photos (e.g. faces or bedrooms). However, these methods are of limited utility because it is difficult for a user to contr ol what the network produces. In this paper, we propose a deep adverserial image synthesis architecture that is conditioned on coarse sketches and sparse color strokes to generate realistic cars, bedrooms, or faces. We demonstrate a sketch based image synthesis system which allows users to 'scribble' over the sketch to indicate preferred color for objects. Our network can then generate convincing images that satisfy both the color and the sketch constraints of user. The netwo rk is feed-forward which allows users to see the effect of their edits in real t ime. We compare to recent work on sketch to image synthesis and show that our ap proach can generate more realistic, more diverse, and more controllable outputs. The architecture is also effective at user-guided colorization of grayscale ima ges.
**********

Scene Graph Generation by Iterative Message Passing
Danfei Xu, Yuke Zhu, Christopher B. Choy, Li Fei-Fei; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5410-5419
Understanding a visual scene goes beyond recognizing individual objects in isola tion. Relationships between objects also constitute rich semantic information ab out the scene. In this work, we explicitly model the objects and their relations hips using scene graphs, a visually-grounded graphical structure of an image. We propose a novel end-to-end model that generates such structured scene represent ation from an input image. Our key insight is that the graph generation problem can be formulated as message passing between the primal node graph and its dual edge graph. Our joint inference model can take advantage of contextual cues to m ake better predictions on objects and their relationships. The experiments show that our model significantly outperforms previous methods on the Visual Genome d ataset as well as support relation inference in NYU Depth V2 dataset.
**********

Accurate Single Stage Detector Using Recurrent Rolling Convolution
Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, Li Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition (CVPR), 2017, pp. 5420-5428
Most of the recent successful methods in accurate object detection and localizat ion used some variants of R-CNN style two stage Convolutional Neural Networks (C NN) where plausible regions were proposed in the first stage then followed by a second stage for decision refinement. Despite the simplicity of training and the efficiency in deployment, the single stage detection methods have not been as c ompetitive when evaluated in benchmarks consider mAP for high IoU thresholds. In this paper, we proposed a novel single stage end-to-end trainable object detect ion network to overcome this limitation. We achieved this by introducing Recurre nt Rolling Convolution (RRC) architecture over multi-scale feature maps to const ruct object classifiers and bounding box regressors which are "deep in context". We evaluated our method in the challenging KITTI dataset which measures methods under IoU threshold of 0.7. We showed that with RRC, a single reduced VGG-16 ba sed model already significantly outperformed all the previously published result

s. At the time this paper was written our models ranked the first in KITTI car d etection (the hard level), the first in cyclist detection and the second in pede strian detection. These results were not reached by the previous single stage me thods. The code is publicly available.
**********

Indoor Scene Parsing With Instance Segmentation, Semantic Labeling and Support R elationship Inference
Wei Zhuo, Mathieu Salzmann, Xuming He, Miaomiao Liu; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5429-5437
Over the years, indoor scene parsing has attracted a growing interest in the com puter vision community. Existing methods have typically focused on diverse subta sks of this challenging problem. In particular, while some of them aim at segmen ting the image into regions, such as object or surface instances, others aim at inferring the semantic labels of given regions, or their support relationships. These different tasks are typically treated as separate ones. However, they bear strong connections: good regions should respect the semantic labels; support ca n only be defined for meaningful regions; support relationships strongly depend on semantics. In this paper, we therefore introduce an approach to jointly segme nt the instances and infer their semantic labels and support relationships from a single input image. By exploiting a hierarchical segmentation, we formulate ou r problem as that of jointly finding the regions in the hierarchy that correspo nd to instances and estimating their class labels and pairwise support relations hips. We express this via a Markov Random Field, which allows us to further enco de links between the different types of variables. Inference in this model can b e done exactly via integer linear programming, and we learn its parameters in a structural SVM framework. Our experiments on NYUv2 demonstrate the benefits of r easoning jointly about all these subtasks of indoor scene parsing.
**********

Reflection Removal Using Low-Rank Matrix Completion
Byeong-Ju Han, Jae-Young Sim; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2017, pp. 5438-5446
The images taken through glass often capture a target transmitted scene as well as undesired reflected scenes. In this paper, we propose a low-rank matrix compl etion algorithm to remove reflection artifacts automatically from multiple glass images taken at slightly different camera locations. We assume that the transmi tted scenes are more dominant than the reflected scenes in typical glass images. We first warp the multiple glass images to a reference image, where the gradien ts are consistent in the transmission images while the gradients are varying acr oss the reflection images. Based on this observation, we compute a gradient reli ability such that the pixels belonging to the salient edges of the transmission image are assigned high reliability. Then we suppress the gradients of the refle ction images and recover the gradients of the transmission images only, by solvi ng a low-rank matrix completion problem in gradient domain. We reconstruct an or iginal transmission image using the resulting optimal gradient map. Experimental results show that the proposed algorithm removes the reflection artifacts from glass images faithfully and outperforms the existing algorithms on typical glass images.
**********

Deep Multimodal Representation Learning From Temporal Data
Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal , Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition (CVPR), 2017, pp. 5447-5455
In recent years, Deep Learning has been successfully applied to multimodal learn ing problems, with the aim of learning useful joint representations in data fusi on applications. When the available modalities consist of time series data such as video, audio and sensor signals, it becomes imperative to consider their temp oral structure during the fusion process. In this paper, we propose the Correlat ional Recurrent Neural Network (CorrRNN), a novel temporal fusion model for fusi ng multiple input modalities that are inherently temporal in nature. Key feature s of our proposed model include: (i) simultaneous learning of the joint represen

tation and temporal dependencies between modalities, (ii) use of multiple loss t erms in the objective function, including a maximum correlation loss term to enh ance learning of cross-modal information, and (iii) the use of an attention mode l to dynamically adjust the contribution of different input modalities to the jo int representation. We validate our model via experimentation on two different t asks: video- and sensor-based activity classification, and audio-visual speech r ecognition. We empirically analyze the contributions of different components of the proposed CorrRNN model, and demonstrate its robustness, effectiveness and st ate-of-the-art performance on multiple datasets.
**********

Weighted-Entropy-Based Quantization for Deep Neural Networks
Eunhyeok Park, Junwhan Ahn, Sungjoo Yoo; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2017, pp. 5456-5464
Quantization is considered as one of the most effective methods to optimize the inference cost of neural network models for their deployment to mobile and embed ded systems, which have tight resource constraints. In such approaches, it is cr itical to provide low-cost quantization under a tight accuracy loss constraint ( e.g., 1%). In this paper, we propose a novel method for quantizing weights and a ctivations based on the concept of weighted entropy. Unlike recent work on binar y-weight neural networks, our approach is multi-bit quantization, in which weigh ts and activations can be quantized by any number of bits depending on the targe t accuracy. This facilitates much more flexible exploitation of accuracy-perform ance trade-off provided by different levels of quantization. Moreover, our schem e provides an automated quantization flow based on conventional training algorit hms, which greatly reduces the design-time effort to quantize the network. Accor ding to our extensive evaluations based on practical neural network models for i mage classification (AlexNet, GoogLeNet and ResNet-50/101), object detection (R- FCN with 50-layer ResNet), and language modeling (an LSTM network), our method a chieves significant reductions in both the model size and the amount of computat ion with minimal accuracy loss. Also, compared to existing quantization schemes, ours provides higher accuracy with a similar resource constraint and requires m uch lower design effort.
**********

Deep Supervision With Shape Concepts for Occlusion-Aware 3D Object Parsing
Chi Li, M. Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D. Hager, Manmohan Chan draker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogn ition (CVPR), 2017, pp. 5465-5474
Monocular 3D object parsing is highly desirable in various scenarios including o cclusion reasoning and holistic scene interpretation. We present a deep convolut ional neural network (CNN) architecture to localize semantic parts in 2D image a nd 3D space while inferring their visibility states, given a single RGB image. O ur key insight is to exploit domain knowledge to regularize the network by deepl y supervising its hidden layers, in order to sequentially infer intermediate con cepts associated with the final task. To acquire training data in desired quanti ties with ground truth 3D shape and relevant concepts, we render 3D object CAD m odels to generate large-scale synthetic data and simulate challenging occlusion configurations between objects. We train the network only on synthetic data and demonstrate state-of-the-art performances on real image benchmarks including an extended version of KITTI, PASCAL VOC, PASCAL3D+ and IKEA for 2D and 3D keypoint localization and instance segmentation. The empirical results substantiate the utility of our deep supervision scheme by demonstrating effective transfer of kn owledge from synthetic data to real images, resulting in less overfitting compar ed to standard end-to-end training.
**********

Spatio-Temporal Self-Organizing Map Deep Network for Dynamic Object Detection Fr om Videos
Yang Du, Chunfeng Yuan, Bing Li, Weiming Hu, Stephen Maybank; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 54 75-5484
In dynamic object detection, it is challenging to construct an effective model t

o sufficiently characterize the spatial-temporal properties of the background. T
his paper proposes a new Spatio-Temporal Self-Organizing Map (STSOM) deep networ
k to detect dynamic objects in complex scenarios. The proposed approach has seve
ral contributions: First, a novel STSOM shared by all pixels in a video frame is
 presented to efficiently model complex background. We exploit the fact that the
 motions of complex background have the global variation in the space and the lo
cal variation in the time, to train STSOM using the whole frames and the sequenc
e of a pixel over time to tackle the variance of complex background. Second, a B
ayesian parameter estimation based method is presented to learn thresholds autom
atically for all pixels to filter out the background. Last, in order to model th
e complex background more accurately, we extend the single-layer STSOM to the de
ep network. Then the background is filtered out layer by layer. Experimental res
ults on  CDnet 2014 dataset demonstrate that the proposed STSOM deep network out
performs numerous recently proposed methods in the overall performance and in mo
st categories of scenarios.
**********

Semantic Image Inpainting With Deep Generative Models
Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Jo
hnson, Minh N. Do; Proceedings of the IEEE Conference on Computer Vision and Pat
tern Recognition (CVPR), 2017, pp. 5485-5493
Semantic image inpainting is a challenging task where large missing regions have
 to be filled based on the available visual data. Existing methods which extract
 information from only a single image generally produce unsatisfactory results d
ue to the lack of high level context.  In this paper, we propose a novel method
for  semantic image inpainting, which generates the missing content by condition
ing on the available data. Given a trained generative model, we search for the c
losest encoding of the corrupted image in the latent image manifold using our co
ntext and prior losses. This encoding is then passed through the generative mode
l to infer the missing content. In our method, inference is possible irrespectiv
e of how the missing content is structured, while the state-of-the-art learning
based method requires  specific information about the holes in the training phas
e. Experiments on three datasets show that our method successfully predicts info
rmation in large missing regions and achieves pixel-level photorealism, signific
antly outperforming the state-of-the-art methods.
**********

Unambiguous Text Localization and Retrieval for Cluttered Scenes
Xuejian Rong, Chucai Yi, Yingli Tian; Proceedings of the IEEE Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2017, pp. 5494-5502
Text instance as one category of self-described objects provides valuable inform
ation for understanding and describing cluttered scenes. In this paper, we explo
re the task of unambiguous text localization and retrieval, to accurately locali
ze a specific targeted text instance in a cluttered image given a natural langua
ge description that refers to it. To address this issue, first a novel recurrent
 Dense Text Localization Network (DTLN) is proposed to sequentially decode the i
ntermediate convolutional representations of a cluttered scene image into a set
of distinct text instance detections. Our approach avoids repeated detections at
 multiple scales of the same text instance by recurrently memorizing previous de
tections, and effectively tackles crowded text instances in close proximity. Sec
ond, we propose a Context Reasoning Text Retrieval (CRTR) model, which jointly e
ncodes text instances and their context information through a recurrent network,
 and ranks localized text bounding boxes by a scoring function of context compat
ibility. Quantitative evaluations on standard scene text localization benchmarks
 and a newly collected scene text retrieval dataset demonstrate the effectivenes
s and advantages of our models for both scene text localization and retrieval.
**********

GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue
Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle,
 Aaron Courville; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2017, pp. 5503-5512
We introduce GuessWhat?!, a two-player guessing game as a testbed for research o

n the interplay of computer vision and dialogue systems. The goal of the game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the proposed task. Our key contribution is the collection of a large-scale dataset consisting of 150K human-played games with a total of 800K visual question-answer pairs on 66K images. We explain our design decisions in collecting the dataset and introduce the oracle and questioner tasks that are associated with the two players of the game. We prototyped deep learning models to establish initial baselines of the introduced tasks.
**********

Learning Spatial Regularization With Image-Level Supervisions for Multi-Label Image Classification
Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5513-5522
Multi-label image classification is a fundamental but challenging task in computer vision. Great progress has been achieved by exploiting semantic relations between labels in recent years. However, conventional approaches are unable to model the underlying spatial relations between labels in multi-label images, because spatial annotations of the labels are generally not provided. In this paper, we propose a unified deep neural network that exploits both semantic and spatial relations between labels with only image-level supervisions. Given a multi-label image, our proposed Spatial Regularization Network (SRN) generates attention maps for all labels and captures the underlying relations between them via learnable convolutions. By aggregating the regularized classification results with original results by a ResNet-101 network, the classification performance can be consistently improved. The whole deep neural network is trained end-to-end with only image-level annotations, thus requires no additional efforts on image annotations. Extensive evaluations on 3 public datasets with different types of labels show that our approach significantly outperforms state-of-the-arts and has strong generalization capability. Analysis of the learned SRN model demonstrates that it can effectively capture both semantic and spatial relations of labels for improving classification performance.
**********

CERN: Confidence-Energy Recurrent Network for Group Activity Recognition
Tianmin Shu, Sinisa Todorovic, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5523-5531
This work is about recognizing human activities occurring in videos at distinct semantic levels, including individual actions, interactions, and group activities. The recognition is realized using a two-level hierarchy of Long Short-Term Memory (LSTM) networks, forming a feed-forward deep architecture, which can be trained end-to-end. In comparison with existing architectures of LSTMs, we make two key contributions giving the name to our approach as Confidence-Energy Recurrent Network -- CERN. First, instead of using the common softmax layer for prediction, we specify a novel energy layer (EL) for estimating the energy of our predictions. Second, rather than finding the common minimum-energy class assignment, which may be numerically unstable under uncertainty, we specify that the EL additionally computes the p-values of the solutions, and in this way estimates the most confident energy minimum. The evaluation on the Collective Activity and Volleyball datasets demonstrates: (i) advantages of our two contributions relative to the common softmax and energy-minimization formulations and (ii) a superior performance relative to the state-of-the-art approaches.
**********

Visual Translation Embedding Network for Visual Relation Detection
Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, Tat-Seng Chua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5532-5540
Visual relations, such as "person ride bike" and "bike next to car", offer a comprehensive scene understanding of an image, and have already shown their great utility in connecting computer vision and natural language. However, due to the c

hallenging combinatorial complexity of modeling subject-predicate-object relatio n triplets, very little work has been done to localize and predict visual relati ons. Inspired by the recent advances in relational representation learning of kn owledge bases and convolutional object detection networks, we propose a Visual T ranslation Embedding network (VTransE) for visual relation detection. VTransE pl aces objects in a low-dimensional relation space where a relation can be modeled as a simple vector translation, i.e., subject + predicate = object. We propose a novel feature extraction layer that enables object-relation knowledge transfer in a fully-convolutional fashion that supports training and inference in a sing le forward/backward pass. To the best of our knowledge, VTransE is the first end -to-end relation detection network. We demonstrate the effectiveness of VTransE over other state-of-the-art methods on two large-scale datasets: Visual Relation ship and Visual Genome. Note that even though VTransE is a purely visual model, it is still competitive to the Lu's multi-modal model with language

**********

Neural Face Editing With Intrinsic Image Disentangling

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, Dimitris Samaras; Proceedings of the IEEE Conference on Computer Vision and Pattern Reco gnition (CVPR), 2017, pp. 5541-5550

Traditional face editing methods often require a number of sophisticated and tas k specific algorithms to be applied one after the other --- a process that is te dious, fragile, and computationally intensive. In this paper, we propose an end- to-end generative adversarial network that infers a face-specific disentangled r epresentation of intrinsic face properties, including shape (i.e. normals), albe do, and lighting, and an alpha matte. We show that this network can be trained o n "in-the-wild" images by incorporating an in-network physically-based image for mation module and appropriate loss functions. Our disentangling latent represent ation allows for semantically relevant edits, where one aspect of facial appeara nce can be manipulated while keeping orthogonal properties fixed, and we demonst rate its use for a number of facial editing applications.

**********

EAST: An Efficient and Accurate Scene Text Detector

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, Jiajun Liang ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5551-5560

Previous approaches for scene text detection have already achieved promising per formances across various benchmarks. However, they usually fall short when deali ng with challenging scenarios, even when equipped with deep neural network model s, because the overall performance is determined by the interplay of multiple st ages and components in the pipelines. In this work, we propose a simple yet powe rful pipeline that yields fast and accurate text detection in natural scenes. Th e pipeline directly predicts words or text lines of arbitrary orientations and q uadrilateral shapes in full images, eliminating unnecessary intermediate steps ( e.g., candidate aggregation and word partitioning), with a single neural network . The simplicity of our pipeline allows concentrating efforts on designing loss functions and neural network architecture. Experiments on standard datasets incl uding ICDAR 2015, COCO-Text and MSRA-TD500 demonstrate that the proposed algorit hm significantly outperforms state-of-the-art methods in terms of both accuracy and efficiency. On the ICDAR 2015 dataset, the proposed algorithm achieves an F- score of 0.7820 at 13.2fps at 720p resolution.

**********

Episodic CAMN: Contextual Attention-Based Memory Networks With Iterative Feedbac k for Scene Labeling

Abrar H. Abdulnabi, Bing Shuai, Stefan Winkler, Gang Wang; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5561- 5570

Scene labeling can be seen as a sequence-sequence prediction task (pixels-labels ), and it is quite important to leverage relevant context to enhance the perform ance of pixel classification. In this paper, we introduce an episodic attention- based memory network to achieve the goal. We present a unified framework that ma

inly consists of a Convolutional Neural Network (CNN), specifically, Fully Convo lutional Network (FCN) and an attention-based memory module with feedback connec tions to perform context selection and refinement. The full model produces conte xt-aware representation for each target patch by aggregating the activated conte xt and its original local representation produced by the convolution layers. We evaluate our model on PASCAL Context, SIFT Flow and PASCAL VOC 2011 datasets and achieve competitive results to other state-of-the-art methods in scene labeling .

**********

Robust Energy Minimization for BRDF-Invariant Shape From Light Fields

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Manmohan Chandraker; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5571-5579

Highly effective optimization frameworks have been developed for traditional mul tiview stereo relying on lambertian photoconsistency. However, they do not accou nt for complex material properties. On the other hand, recent works have explore d PDE invariants for shape recovery with complex BRDFs, but they have not been i ncorporated into robust numerical optimization frameworks. We present a variatio nal energy minimization framework for robust recovery of shape in multiview ster eo with complex, unknown BRDFs. While our formulation is general, we demonstrate its efficacy on shape recovery using a single light field image, where the micr olens array may be considered as a realization of a purely translational multivi ew stereo setup. Our formulation automatically balances contributions from textu re gradients, traditional Lambertian photoconsistency, an appropriate BRDF-invar iant PDE and a smoothness prior. Unlike prior works, our energy function inheren tly handles spatially-varying BRDFs and albedos. Extensive experiments with synt hetic and real data show that our optimization framework consistently achieves e rrors lower than Lambertian baselines and further, is more robust than prior BRD F-invariant reconstruction methods.

**********

Connecting Look and Feel: Associating the Visual and Tactile Properties of Physi cal Materials

Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, Edward Adelson; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5580-5588

For machines to interact with the physical world, they must understand the physi cal properties of objects and materials they encounter. We use fabrics as an exa mple of a deformable material with a rich set of mechanical properties. A thin f lexible fabric, when draped, tends to look different from a heavy stiff fabric. It also feels different when touched. Using a collection of 118 fabric samples, we captured color and depth images of draped fabrics along with tactile data fro m a high-resolution touch sensor. We then sought to associate the information fr om vision and touch by jointly training CNN's across the three modalities. Throu gh the CNN, each input, regardless of the modality, generates an embedding vecto r that records the fabric's physical property. By comparing the embedding vector s, our system is able to look at a fabric image and predict how it will feel, an d vice versa. We also show that a system jointly trained on vision and touch dat a can outperform a similar system trained only on visual data when tested purely with visual inputs.

**********

Robust Visual Tracking Using Oblique Random Forests

Le Zhang, Jagannadan Varadarajan, Ponnuthurai Nagaratnam Suganthan, Narendra Ahu ja, Pierre Moulin; Proceedings of the IEEE Conference on Computer Vision and Pat tern Recognition (CVPR), 2017, pp. 5589-5598

Random forest has emerged as a powerful classification technique with promising results in various vision tasks including image classification, pose estimation and object detection. However, current techniques have shown little improvements in visual tracking as they mostly rely on piece wise orthogonal hyperplanes to create decision nodes and lack a robust incremental learning mechanism that is m uch needed for online tracking. In this paper, we propose a discriminative track

er based on a novel incremental oblique random forest. Unlike conventional ortho gonal decision trees that use a single feature and heuristic measures to obtain a split at each node, we propose to use a more powerful proximal SVM to obtain o blique hyperplanes to capture the geometric structure of the data better. The re sulting decision surface is not restricted to be axis aligned and hence has the ability to represent and classify the input data better. Furthermore, in order t o generalize to online tracking scenarios, we derive incremental update steps th at enable the hyperplanes in each node to be updated recursively, efficiently an d in a closed-form fashion. We demonstrate the effectiveness of our method using two large scale benchmark datasets (OTB-51 and OTB-100) and show that our metho d gives competitive results on several challenging cases by relying on simple HO G features as well as in combination with more sophisticated deep neural network based models. The implementations of the proposed random forest are available a t https://github.com/ZhangLeUestc/ Incremental-Oblique-Random-Forest.

**********

Discriminative Covariance Oriented Representation Learning for Face Recognition With Image Sets

Wen Wang, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5599-5608

For face recognition with image sets, while most existing works mainly focus on building robust set models with hand-crafted feature, it remains a research gap to learn better image representations which can closely match the subsequent ima ge set modeling and classification. Taking sample covariance matrix as set model in the light of its recent promising success, we present a Discriminative Covar iance oriented Representation Learning (DCRL) framework to bridge the above gap. The framework constructs a feature learning network (e.g. a CNN) to project the face images into a target representation space, and the network is trained towa rds the goal that the set covariance matrix calculated in the target space has m aximum discriminative ability. To encode the discriminative ability of set covar iance matrices, we elaborately design two different loss functions, which respec tively lead to two different representation learning schemes, i.e., the Graph Em bedding scheme and the Softmax Regression scheme. Both schemes optimize the whol e network containing both image representation mapping and set model classificat ion in a joint learning manner. The proposed method is extensively validated on three challenging and large scale databases for the task of face recognition wit h image sets, i.e., YouTube Celebrities, YouTube Face DB and Point-and-Shoot Cha llenge.

**********

Generalized Deep Image to Image Regression

Venkataraman Santhanam, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5609-56 19

We present a Deep Convolutional Neural Network architecture which serves as a ge neric image-to-image regressor that can be trained end-to-end without any furthe r machinery. Our proposed architecture, the Recursively Branched Deconvolutional Network (RBDN), develops a cheap multi-context image representation very early on using an efficient recursive branching scheme with extensive parameter sharin g and learnable upsampling. This multi-context representation is subjected to a highly non-linear locality preserving transformation by the remainder of our net work comprising of a series of convolutions/deconvolutions without any spatial d ownsampling. The RBDN architecture is fully convolutional and can handle variabl e sized images during inference. We provide qualitative/quantitative results on 3 diverse tasks: relighting, denoising and colorization and show that our propos ed RBDN architecture obtains comparable results to the state-of-the-art on each of these tasks when used off-the-shelf without any post processing or task-speci fic architectural modifications.

**********

Multi-Object Tracking With Quadruplet Convolutional Neural Networks

Jeany Son, Mooyeol Baek, Minsu Cho, Bohyung Han; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5620-5629

We propose Quadruplet Convolutional Neural Networks (Quad-CNN) for multi-object tracking, which learn to associate object detections across frames using quadruplet losses. The proposed networks consider target appearances together with their temporal adjacencies for data association. Unlike conventional ranking losses, the quadruplet loss enforces an additional constraint that makes temporally adjacent detections more closely located than the ones with large temporal gaps. We also employ a multi-task loss to jointly learn object association and bounding box regression for better localization. The whole network is trained end-to-end. For tracking, the target association is performed by minimax label propagation using the metric learned from the proposed network. We evaluate performance of our multi-object tracking algorithm on public MOT Challenge datasets, and achieve outstanding results.

**********

Semantic Compositional Networks for Visual Captioning

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, Li Deng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5630-5639

A Semantic Compositional Network (SCN) is developed for image captioning, in which semantic concepts (i.e., tags) are detected from the image, and the probability of each tag is used to compose the parameters in a long short-term memory (LSTM) network. The SCN extends each weight matrix of the LSTM to an ensemble of tag-dependent weight matrices. The degree to which each member of the ensemble is used to generate an image caption is tied to the image-dependent probability of the corresponding tag. In addition to captioning images, we also extend the SCN to generate captions for video clips. We qualitatively analyze semantic composition in SCNs, and quantitatively evaluate the algorithm on three benchmark datasets: COCO, Flickr30k, and Youtube2Text. Experimental results show that the proposed method significantly outperforms prior state-of-the-art approaches, across multiple evaluation metrics.

**********

Link the Head to the "Beak": Zero Shot Learning From Noisy Text Description at Part Precision

Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, Ahmed Elgammal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5640-5649

In this paper, we study learning visual classifiers from unstructured text description at part precision with no training images. We show that visual text terms can be encouraged to attend to its relevant parts, while image connections to non-visual text terms vanishes without any supervision. This learning process enables terms like "peak" to be linked to parts like only head for instance , while non-visual terms like "migrate" not to affect classifier prediction without part-text annotation. Images are encoded by a part-based CNN that detect bird parts and learn part-specific learning representation. Part-based visual classifiers are predicted from text descriptions of unseen visual classifiers to facilitate classification without training images (also known as zero-shot recognition ). We performed our experiments on CUB200 dataset and improves the zero-shot recognition results from 34.2% to 44.0%. We also created a large scale benchmark on 404 North American Bird Images with text descriptions, where we also showed that our method outperforming existing methods.

**********

A Non-Local Low-Rank Framework for Ultrasound Speckle Reduction

Lei Zhu, Chi-Wing Fu, Michael S. Brown, Pheng-Ann Heng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5650-5658

`Speckle' refers to the granular patterns that occur in ultrasound images due to wave interference. Speckle removal can greatly improve the visibility of the underlying structures in an ultrasound image and enhance subsequent post processing. We present a novel framework for speckle removal based on low-rank non-local filtering. Our approach works by first computing a guidance image that assists in the selection of candidate patches for non-local filtering in the face of sig

nificant speckles. The candidate patches are further refined using a low-rank m inimization estimated using a truncated weighted nuclear norm (TWNN) and structu red sparsity. We show that the proposed filtering framework produces results th at outperform state-of-the-art methods both qualitatively and quantitatively. T his framework also provides better segmentation results when used for pre-proces sing ultrasound images.
**********

SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning
Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Ch ua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio n (CVPR), 2017, pp. 5659-5667
Visual attention has been successfully applied in structural prediction tasks su ch as visual captioning and question answering. Existing visual attention models are generally spatial, i.e., the attention is modeled as spatial probabilities that re-weight the last conv-layer feature map of a CNN encoding an input image. However, we argue that such spatial attention does not necessarily conform to t he attention mechanism --- a dynamic feature extractor that combines contextual fixations over time, as CNN features are naturally spatial, channel-wise and mul ti-layer. In this paper, we introduce a novel convolutional neural network dubbe d SCA-CNN that incorporates Spatial and Channel-wise Attentions in a CNN. In the task of image captioning, SCA-CNN dynamically modulates the sentence generation context in multi-layer feature maps, encoding where (i.e., attentive spatial lo cations at multiple layers) and what (i.e., attentive channels) the visual atten tion is. We evaluate the proposed SCA-CNN architecture on three benchmark image captioning datasets: Flickr8K, Flickr30K, and MSCOCO. It is consistently observe d that SCA-CNN significantly outperforms state-of-the-art visual attention-based image captioning methods.
**********

A Compact DNN: Approaching GoogLeNet-Level Accuracy of Classification and Domain Adaptation
Chunpeng Wu, Wei Wen, Tariq Afzal, Yongmei Zhang, Yiran Chen, Hai (Helen) Li; Pr oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP R), 2017, pp. 5668-5677
Recently, DNN model compression based on network architecture design, e.g., Sque ezeNet, attracted a lot attention. No accuracy drop on image classification is o bserved on these extremely compact networks, compared to well-known models. An e merging question, however, is whether these model compression techniques hurt DN N's learning ability other than classifying images on a single dataset. Our prel iminary experiment shows that these compression methods could degrade domain ada ptation (DA) ability, though the classification performance is preserved. Theref ore, we propose a new compact network architecture and unsupervised DA method in this paper. The DNN is built on a new basic module Conv-M which provides more d iverse feature extractors without significantly increasing parameters. The unifi ed framework of our DA method will simultaneously learn invariance across domain s, reduce divergence of feature representations, and adapt label prediction. Our DNN has 4.1M parameters, which is only 6.7% of AlexNet or 59% of GoogLeNet. Exp eriments show that our DNN obtains GoogLeNet-level accuracy both on classificati on and DA, and our DA method slightly outperforms previous competitive ones. Put all together, our DA strategy based on our DNN achieves state-of-the-art on six teen of total eighteen DA tasks on popular Office-31 and Office-Caltech datasets .
**********

Relationship Proposal Networks
Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, Ahmed Elgammal; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 17, pp. 5678-5686
Image scene understanding requires learning the relationships between objects in the scene. A scene with many objects may have only a few individual interacting objects (e.g., in a party image with many people, only a handful of people migh

t be speaking with each other). To detect all relationships, it would be ineffi cient to first detect all individual objects and then classify all pairs; not on ly is the number of all pairs quadratic, but classification requires limited obj ect categories, which is not scalable for real-world images. In this paper we ad dress these challenges by using pairs of related regions in images to train a re lationship proposer that at test time produces a manageable number of related re gions. We name our model the Relationship Proposal Network (Rel-PN). Like object proposals, our Rel-PN is class-agnostic and thus scalable to an open vocabulary of objects. We demonstrate the ability of our Rel-PN to localize relationships with only a few thousand proposals. We demonstrate its performance on the Visua l Genome dataset and compare to other baselines that we designed. We also conduc t experiments on a smaller subset of 5,000 images with over 37,000 related regio ns and show promising results.

**********

Designing Energy-Efficient Convolutional Neural Networks Using Energy-Aware Prun ing

Tien-Ju Yang, Yu-Hsin Chen, Vivienne Sze; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5687-5695

Deep convolutional neural networks (CNNs) are indispensable to state-of-the-art computer vision algorithms. However, they are still rarely deployed on battery-p owered mobile devices, such as smartphones and wearable gadgets, where vision al gorithms can enable many revolutionary real-world applications. The key limiting factor is the high energy consumption of CNN processing due to its high computa tional complexity. While there are many previous efforts that try to reduce the CNN model size or the amount of computation, we find that they do not necessaril y result in lower energy consumption. Therefore, these targets do not serve as a good metric for energy cost estimation. To close the gap between CNN design an d energy consumption optimization, we propose an energy-aware pruning algorithm for CNNs that directly uses the energy consumption of a CNN to guide the pruning process. The energy estimation methodology uses parameters extrapolated from ac tual hardware measurements. The proposed layer-by-layer pruning algorithm also p runes more aggressively than previously proposed pruning methods by minimizing t he error in the output feature maps instead of the filter weights. For each laye r, the weights are first pruned and then locally fine-tuned with a closed-form l east-square solution to quickly restore the accuracy. After all layers are prune d, the entire network is globally fine-tuned using back-propagation. With the pr oposed pruning method, the energy consumption of AlexNet and GoogLeNet is reduce d by 3.7x and 1.6x, respectively, with less than 1% top-5 accuracy loss. We also show that reducing the number of target classes in AlexNet greatly decreases th e number of weights, but has a limited impact on energy consumption.

**********

Boundary-Aware Instance Segmentation

Zeeshan Hayder, Xuming He, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5696-5704

We address the problem of instance-level semantic seg- mentation, which aims at jointly detecting, segmenting and classifying every individual object in an imag e. In this con- text, existing methods typically propose candidate objects, usua lly as bounding boxes, and directly predict a binary mask within each such propo sal. As a consequence, they cannot recover from errors in the object candidate g enera- tion process, such as too small or shifted boxes. In this paper, we intro duce a novel object segment rep- resentation based on the distance transform of the object masks. We then design an object mask network (OMN) with a new residua l-deconvolution architecture that infers such a representation and decodes it in to the final binary object mask. This allows us to predict masks that go beyond the scope of the bounding boxes and are thus robust to inaccu- rate object candi dates. We integrate our OMN into a Mul- titask Network Cascade framework, and le arn the result- ing boundary-aware instance segmentation (BAIS) network in an en d-to-end manner. Our experiments on the PAS- CAL VOC 2012 and the Cityscapes dat asets demonstrate the benefits of our approach, which outperforms the state- of-the-art in both object proposal generation and instance segmentation.

**********
Joint Intensity and Spatial Metric Learning for Robust Gait Recognition
Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5705-5715
This paper describes a joint intensity metric learning method to improve the robustness of gait recognition with silhouette-based descriptors such as gait energy images. Because existing methods often use the difference of image intensities between a matching pair (e.g., the absolute difference of gait energies for the l_1-norm) to measure a dissimilarity, large intrasubject differences derived from covariate conditions (e.g., large gait energies caused by carried objects vs. small gait energies caused by the background), may wash out subtle intersubject differences (e.g., the difference of middle-level gait energies derived from motion differences). We therefore introduce a metric on joint intensity to mitigate the large intrasubject differences as well as leverage the subtle intersubject differences. More specifically, we formulate the joint intensity and spatial metric learning in a unified framework and alternately optimize it by linear or ranking support vector machines. Experiments using the OU-ISIR treadmill data set B with the largest clothing variation and large population data set with bag, b version containing carrying status in the wild demonstrate the effectiveness of the proposed method.
**********
Seeing What Is Not There: Learning Context to Determine Where Objects Are Missing
Jin Sun, David W. Jacobs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5716-5724
Most of computer vision focuses on what is in an image. We propose to train a standalone object-centric context representation to perform the opposite task: seeing what is not there. Given an image, our context model can predict where objects should exist, even when no object instances are present. Combined with object detection results, we can perform a novel vision task: finding where objects are missing in an image. Our model is based on a convolutional neural network structure. With a specially designed training strategy, the model learns to ignore objects and focus on context only. It is fully convolutional thus highly efficient. Experiments show the effectiveness of the proposed approach in one important accessibility task: finding city street regions where curb ramps are missing, which could help millions of people with mobility disabilities.
**********
Joint Gap Detection and Inpainting of Line Drawings
Kazuma Sasaki, Satoshi Iizuka, Edgar Simo-Serra, Hiroshi Ishikawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5725-5733
We propose a novel data-driven approach for automatically detecting and completing gaps in line drawings with a Convolutional Neural Network. In the case of existing inpainting approaches for natural images, masks indicating the missing regions are generally required as input. Here, we show that line drawings have enough structures that can be learned by the CNN to allow automatic detection and completion of the gaps without any such input. Thus, our method can find the gaps in line drawings and complete them without user interaction. Furthermore, the completion realistically conserves thickness and curvature of the line segments. All the necessary heuristics for such realistic line completion are learned naturally from a dataset of line drawings, where various patterns of line completion are generated on the fly as training pairs to improve the model generalization. We evaluate our method qualitatively on a diverse set of challenging line drawings and also provide quantitative results with a user study, where it significantly outperforms the state of the art.
**********
CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos
Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, Shih-Fu Chang; Pr

oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP
R), 2017, pp. 5734-5743
Temporal action localization is an important yet challenging problem. Given a lo
ng, untrimmed video consisting of multiple action instances and complex backgrou
nd contents, we need not only to recognize their action categories, but also to
localize the start time and end time of each instance. Many state-of-the-art sys
tems use segment-level classifiers to select and rank proposal segments of pre-d
etermined boundaries. However, a desirable model should move beyond segment-leve
l and make dense predictions at a fine granularity in time to determine precise
temporal boundaries. To this end, we design a novel Convolutional-De-Convolution
al (CDC) network that places CDC filters on top of 3D ConvNets, which have been
shown to be effective for abstracting action semantics but reduce the temporal l
ength of the input data. The proposed CDC filter performs the required temporal
upsampling and spatial downsampling operations simultaneously to predict actions
 at the frame-level granularity. It is unique in jointly modeling action semanti
cs in space-time and fine-grained temporal dynamics. We train the CDC network in
 an end-to-end manner efficiently. Our model not only achieves superior performa
nce in detecting actions in every frame, but also significantly boosts the preci
sion of localizing temporal boundaries. Finally, the CDC network demonstrates a
very high efficiency with the ability to process 500 frames per second on a sing
le GPU server. Source code and trained models are available online at https://bi
tbucket.org/columbiadvmm/cdc.
**********

Switching Convolutional Neural Network for Crowd Counting
Deepak Babu Sam, Shiv Surya, R. Venkatesh Babu; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5744-5752
We propose a novel crowd counting model that maps a given crowd scene to its den
sity. Crowd analysis is compounded by myriad of factors like inter-occlusion bet
ween people due to extreme crowding, high similarity of appearance between peopl
e and background elements, and large variability of camera view-points. Current
state-of-the art approaches tackle these factors by using multi-scale CNN archit
ectures, recurrent networks and late fusion of features from multi-column CNN wi
th different receptive fields. We propose switching convolutional neural network
 that leverages variation of crowd density within an image to improve the accura
cy and localization of the predicted crowd count. Patches from a grid within a c
rowd scene are relayed to independent CNN regressors based on crowd count predic
tion quality of the CNN established during training. The independent CNN regress
ors are designed to have different receptive fields and a switch classifier is t
rained to relay the crowd scene patch to the best CNN regressor. We perform exte
nsive experiments on all major crowd counting datasets and evidence better perfo
rmance compared to current state-of-the-art methods. We provide interpretable re
presentations of the multichotomy of space of crowd scene patches inferred from
the switch. It is observed that the switch relays an image patch to a particular
 CNN column based on density of crowd.
**********

Captioning Images With Diverse Objects
Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Tr
evor Darrell, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2017, pp. 5753-5761
Recent captioning models are limited in their ability to scale and describe conc
epts unseen in  paired image-text corpora. We propose the Novel Object Captioner
 (NOC), a deep visual semantic captioning model that can describe a large number
 of object categories not present in existing image-caption datasets. Our model
takes advantage of external sources -- labeled images from object recognition da
tasets, and semantic knowledge extracted from unannotated text.  We propose mini
mizing a joint objective which can learn from these diverse data sources and lev
erage distributional semantic embeddings,  enabling the model to generalize and
describe novel objects outside of image-caption datasets.  We demonstrate that o
ur model exploits semantic information to generate captions for hundreds of obje
ct categories in the ImageNet object recognition dataset that are not observed i

n MSCOCO image-caption training data, as well as many categories that are observ ed very rarely. Both automatic evaluations and human judgements show that our mo del considerably outperforms prior work in being able to describe many more cate gories of objects.
**********

Amodal Detection of 3D Objects: Inferring 3D Bounding Boxes From 2D Ones in RGB-Depth Images
Zhuo Deng, Longin Jan Latecki; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2017, pp. 5762-5770
This paper addresses the problem of amodal perception of 3D object detection. Th e task is to not only find object localizations in the 3D world, but also estima te their physical sizes and poses, even if only parts of them are visible in the RGB-D image. Recent approaches have attempted to harness point cloud from depth channel to exploit 3D features directly in the 3D space and demonstrated the su periority over traditional 2.5D representation approaches. We revisit the amodal 3D detection problem by sticking to the 2.5D representation framework, and dire ctly relate 2.5D visual appearance to 3D objects. We propose a novel 3D object d etection system that simultaneously predicts objects' 3D locations, physical siz es, and orientations in indoor scenes. Experiments on the NYUV2 dataset show our algorithm significantly outperforms the state-of-the-art and indicates 2.5D rep resentation is capable of encoding features for 3D amodal object detection. All source code and data is on https://github.com/phoenixnn/Amodal3Det.
**********

Consistent-Aware Deep Learning for Person Re-Identification in a Camera Network
Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, Jie Zhou; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5771 -5780
In this paper, we propose a consistent-aware deep learning (CADL) framework for person re-identification in a camera network. Unlike most existing person re-ide ntification methods which identify whether two body images are from the same per son, our approach aims to obtain the maximal correct matches for the whole camer a network. Different from recently proposed camera network based re-identificati on methods which only consider the consistent information in the matching stage to obtain a global optimal association, we exploit such consistent-aware informa tion under a deep learning framework where both feature representation and image matching are automatically learned with certain consistent constraints. Specifi cally, we reach the global optimal solution and balance the performance between different cameras by optimizing the similarity and association iteratively. Expe rimental results show that our method obtains significant performance improvemen t and outperforms the state-of-the-art methods by large margins.
**********

Enhancing Video Summarization via Vision-Language Embedding
Bryan A. Plummer, Matthew Brown, Svetlana Lazebnik; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5781-5789
This paper addresses video summarization, or the problem of distilling a raw vid eo into a shorter form while still capturing the original story. We show that vi sual representations supervised by freeform language make a good fit for this ap plication by extending a recent submodular summarization approach with represent ativeness and interestingness objectives computed on features from a joint visio n-language embedding space. We perform an evaluation on two diverse datasets, UT Egocentric and TV Episodes, and show that our new objectives give improved summ arization ability compared to standard visual features alone. Our experiments al so show that the vision-language embedding need not be trained on domain specifi c data, but can be learned from standard still image vision-language datasets an d transferred to video. A further benefit of our model is the ability to guide a summary using freeform text input at test time, allowing user customization.
**********

Quality Aware Network for Set to Set Recognition
Yu Liu, Junjie Yan, Wanli Ouyang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5790-5799

This paper targets on the problem of set to set recognition, which learns the metric between two image sets.  Images in each set belong to the same identity.  Since images in a set can be complementary, they hopefully lead to higher accuracy in practical applications. However, the quality of each sample cannot be guaranteed, and samples with poor quality will hurt the metric. In this paper, the quality aware network (QAN) is proposed to confront this problem, where the quality of each sample can be automatically learned although such information is not explicitly provided in the training stage. The network has two branches, where the first branch extracts appearance feature embedding for each sample and the other branch predicts quality score for each sample. Features and quality scores of all samples in a set are then aggregated to generate the final feature embedding. We show that the two branches can be trained in an end-to-end manner given only the set-level identity annotation. Analysis on gradient spread of this mechanism indicates that the quality learned by the network is beneficial to set-to-set recognition and simplifies the distribution that the network needs to fit. Experiments on both face verification and person re-identification show advantages of the proposed QAN. The source code and network structure can be downloaded at GitHub.
**********
Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes
S. Alireza Golestaneh, Lina J. Karam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5800-5809
The detection of spatially-varying blur without having any information about the blur type is a challenging task. In this paper, we propose a novel effective approach to address this blur detection problem from a single image without requiring any knowledge about the blur type, level, or camera settings. Our approach computes blur detection maps based on a novel High-frequency multiscale Fusion and Sort Transform (HiFST) of gradient magnitudes.  The evaluations of the proposed approach on a diverse set of blurry images with different blur types, levels, and contents demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods qualitatively and quantitatively.
**********
Age Progression/Regression by Conditional Adversarial Autoencoder
Zhifei Zhang, Yang Song, Hairong Qi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5810-5818
"If I provide you a face image of mine (without telling you the actual age when I took the picture) and a large amount of face images that I crawled (containing labeled faces of different ages but not necessarily paired), can you show me what I would look like when I am 80 or what I was like when I was 5?" The answer is probably a "No." Most existing face aging works attempt to learn the transformation between age groups and thus would require the paired samples as well as the labeled query image. In this paper, we look at the problem from a generative modeling perspective such that no paired samples is required. In addition, given an unlabeled image, the generative model can directly produce the image with desired age attribute. We propose a conditional adversarial autoencoder (CAAE) that learns a face manifold, traversing on which smooth age progression and regression can be realized simultaneously. In CAAE, the face is first mapped to a latent vector through a convolutional encoder, and then the vector is projected to the face manifold conditional on age through a deconvolutional generator. The latent vector preserves personalized face features (i.e., personality) and the age condition controls progression vs. regression. Two adversarial networks are imposed on the encoder and generator, respectively, forcing to generate more photo-realistic faces. Experimental results demonstrate the appealing performance and flexibility of the proposed framework by comparing with the state-of-the-art and ground truth.
**********
Residual Expansion Algorithm: Fast and Effective Optimization for Nonconvex Least Squares Problems
Daiki Ikami, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5819-5827
We propose the residual expansion (RE) algorithm: a global (or near-global) optimization method for nonconvex least squares problems. Unlike most existing nonconvex optimization techniques, the RE algorithm is not based on either stochastic or multi-point searches; therefore, it can achieve fast global optimization. Moreover, the RE algorithm is easy to implement and successful in high-dimensional optimization. The RE algorithm exhibits excellent empirical performance in terms of k-means clustering, point-set registration, optimized product quantization, and blind image deblurring.
**********

ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes
Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, Matthias Niessner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5828-5839
A key requirement for leveraging supervised deep learning methods is the availability of large, labeled datasets. Unfortunately, in the context of RGB-D scene understanding, very little data is available -- current datasets cover a small range of scene views and have limited semantic annotations. To address this issue, we introduce ScanNet, an RGB-D video dataset containing 2.5M views in 1513 scenes annotated with 3D camera poses, surface reconstructions, and semantic segmentations. To collect this data, we designed an easy-to-use and scalable RGB-D capture system that includes automated surface reconstruction and crowdsourced semantic annotation. We show that using this data helps achieve state-of-the-art performance on several 3D scene understanding tasks, including 3D object classification, semantic voxel labeling, and CAD model retrieval.
**********

More Is Less: A More Complicated Network With Less Inference Complexity
Xuanyi Dong, Junshi Huang, Yi Yang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5840-5848
In this paper, we present a novel and general network structure towards accelerating the inference process of convolutional neural networks, which is more complicated in network structure yet with less inference complexity. The core idea is to equip each original convolutional layer with another low-cost collaborative layer (LCCL), and the element-wise multiplication of the ReLU outputs of these two parallel layers produces the layer-wise output. The combined layer is potentially more discriminative than the original convolutional layer, and its inference is faster for two reasons: 1) the zero cells of the LCCL feature maps will remain zero after element-wise multiplication, and thus it is safe to skip the calculation of the corresponding high-cost convolution in the original convolutional layer; 2) LCCL is very fast if it is implemented as a 1*1 convolution or only a single filter shared by all channels. Extensive experiments on the CIFAR-10, CIFAR-100 and ILSCRC-2012 benchmarks show that our proposed network structure can accelerate the inference process by 32% on average with negligible performance drop.
**********

Online Video Object Segmentation via Convolutional Trident Network
Won-Dong Jang, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5849-5858
A semi-supervised online video object segmentation algorithm, which accepts user annotations about a target object at the first frame, is proposed in this work. We propagate the segmentation labels at the previous frame to the current frame using optical flow vectors. However, the propagation is error-prone. Therefore, we develop the convolutional trident network (CTN), which has three decoding branches: separative, definite foreground, and definite background decoders. Then, we perform Markov random field optimization based on outputs of the three decoders. We sequentially carry out these processes from the second to the last frames to extract a segment track of the target object. Experimental results demonstrate that the proposed algorithm significantly outperforms the state-of-the-art conventional algorithms on the DAVIS benchmark dataset.
**********

Learning Object Interactions and Descriptions for Semantic Image Segmentation
Guangrun Wang, Ping Luo, Liang Lin, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5859-5867
Recent advanced deep convolutional networks (CNNs) achieved great successes in many computer vision tasks, because of their compelling learning complexity and the presences of large-scale labeled data. However, as obtaining per-pixel annotations is expensive, performances of CNNs in semantic image segmentation are not fully exploited. This work significantly increases segmentation accuracy of CNNs by learning from an Image Descriptions in the Wild (IDW) dataset. Unlike previous image captioning datasets, where captions were manually and densely annotated, images and their descriptions in IDW are automatically downloaded from Internet without any manual cleaning and refinement. An IDW-CNN is proposed to jointly train IDW and existing image segmentation dataset such as Pascal VOC 2012 (VOC). It has two appealing properties. First, knowledge from different datasets can be fully explored and transferred from each other to improve performance. Second, segmentation accuracy in VOC can be constantly increased when selecting more data from IDW. Extensive experiments demonstrate the effectiveness and scalability of IDW-CNN, which outperforms existing best-performing system by 12% on VOC12 test set.
**********
Shape Completion Using 3D-Encoder-Predictor CNNs and Shape Synthesis
Angela Dai, Charles Ruizhongtai Qi, Matthias Niessner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5868-5877
We introduce a data-driven approach to complete partial 3D shapes through a combination of volumetric deep neural networks and 3D shape synthesis. From a partially-scanned input shape, our method first infers a low-resolution -- but complete -- output. To this end, we introduce a 3D-Encoder-Predictor Network (3D-EPN) which is composed of 3D convolutional layers. The network is trained to predict and fill in missing data, and operates on an implicit surface representation that encodes both known and unknown space. This allows us to predict global structure in unknown areas at high accuracy. We then correlate these intermediary results with 3D geometry from a shape database at test time. In a final pass, we propose a patch-based 3D shape synthesis method that imposes the 3D geometry from these retrieved shapes as constraints on the coarsely-completed mesh. This synthesis process enables us to reconstruct fine-scale detail and generate high-resolution output while respecting the global mesh structure obtained by the 3D-EPN. Although our 3D-EPN outperforms state-of-the-art completion method, the main contribution in our work lies in the combination of a data-driven shape predictor and analytic 3D shape synthesis. In our results, we show extensive evaluations on a newly-introduced shape completion benchmark for both real-world and synthetic data.
**********
The Impact of Typicality for Informative Representative Selection
Jawadul H. Bappy, Sujoy Paul, Ertem Tuncel, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5878-5887
In computer vision, selection of the most informative samples from a huge pool of training data in order to learn a good recognition model is an active research problem. Furthermore, it is also useful to reduce the annotation cost, as it is time consuming to annotate unlabeled samples. In this paper, motivated by the theories in data compression, we propose a novel sample selection strategy which exploits the concept of typicality from the domain of information theory. Typicality is a simple and powerful technique which can be applied to compress the training data to learn a good classification model. In this work, typicality is used to identify a subset of the most informative samples for labeling, which is then used to update the model using active learning. The proposed model can take advantage of the inter-relationships between data samples. Our approach leads to a significant reduction of manual labeling cost while achieving similar or better recognition performance compared to a model trained with entire training set. This is demonstrated through rigorous experimentation on five datasets.

**********
Infinite Variational Autoencoder for Semi-Supervised Learning
M. Ehsan Abbasnejad, Anthony Dick, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5888-5897
This paper presents an infinite variational autoencoder (VAE) whose capacity adapts to suit the input data. This is achieved using a mixture model where the mixing coefficients are modeled by a Dirichlet process, allowing us to integrate over the coefficients when performing inference. Critically, this then allows us to automatically vary the number of autoencoders in the mixture based on the data. Experiments show the flexibility of our method, particularly for semi-supervised learning, where only a small number of training samples are available.
**********
Understanding Traffic Density From Large-Scale Web Camera Data
Shanghang Zhang, Guanhang Wu, Joao P. Costeira, Jose M. F. Moura; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5898-5907
Understanding traffic density from large-scale web camera (webcam) videos is a challenging problem because such videos have low spatial and temporal resolution, high occlusion and large perspective. To deeply understand traffic density, we explore both optimization based and deep learning based methods. To avoid individual vehicle detection or tracking, both methods map the dense image feature into vehicle density, one based on rank constrained regression and the other based on fully convolutional networks (FCN). The regression based method learns different weights for different blocks of the image to embed road geometry and significantly reduce the error induced by camera perspective. The FCN based method jointly estimates vehicle density and vehicle count with a residual learning framework to perform end-to-end dense prediction, allowing arbitrary image resolution, and adapting to different vehicle scales and perspectives. We analyze and compare both methods, and get insights from optimization based method to improve deep model. Since existing datasets do not cover all the challenges in our work, we collected and labelled a large-scale traffic video dataset, containing 60 million frames from 212 webcams. Both methods are extensively evaluated and compared on different counting tasks and datasets. FCN based method significantly reduces the mean absolute error (MAE) from 10.99 to 5.31 on the public dataset TRANCOS compared with the state-of-the-art baseline.
**********
End-To-End 3D Face Reconstruction With Deep Neural Networks
Pengfei Dou, Shishir K. Shah, Ioannis A. Kakadiaris; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5908-5917
Monocular 3D facial shape reconstruction from a single 2D facial image has been an active research area due to its wide applications. Inspired by the success of deep neural networks (DNN), we propose a DNN-based approach for End-to-End 3D FAce Reconstruction (UH-E2FAR) from a single 2D image. Different from recent works that reconstruct and refine the 3D face in an iterative manner using both an RGB image and an initial 3D facial shape rendering, our DNN model is end-to-end, and thus the complicated 3D rendering process can be avoided. Moreover, we integrate in the DNN architecture two components, namely a multi-task loss function and a fusion convolutional neural network (CNN) to improve facial expression reconstruction. With the multi-task loss function, 3D face reconstruction is divided into neutral 3D facial shape reconstruction and expressive 3D facial shape reconstruction. The neutral 3D facial shape is class-specific. Therefore, higher layer features are useful. In comparison, the expressive 3D facial shape favors lower or intermediate layer features. With the fusion-CNN, features from different intermediate layers are fused and transformed for predicting the 3D expressive facial shape. Through extensive experiments, we demonstrate the superiority of our end-to-end framework in improving the accuracy of 3D face reconstruction.
**********
Deep Learning With Low Precision by Half-Wave Gaussian Quantization
Zhaowei Cai, Xiaodong He, Jian Sun, Nuno Vasconcelos; Proceedings of the IEEE Co

nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5918-5926
The problem of quantizing the activations of a deep neural network is considered. An examination of the popular binary quantization approach shows that this consists of approximating a classical non-linearity, the hyperbolic tangent, by two functions: a piecewise constant sign function, which is used in feedforward network computations, and a piecewise linear hard tanh function, used in the backpropagation step during network learning. The problem of approximating the widely used ReLU non-linearity is then considered. An half-wave Gaussian quantizer (HWGQ) is proposed for forward approximation and shown to have efficient implementation, by exploiting the statistics of of network activations and batch normalization operations. To overcome the problem of gradient mismatch, due to the use of different forward and backward approximations, several piece-wise backward approximators are then investigated. The implementation of the resulting quantized network, denoted as HWGQ-Net, is shown to achieve much closer performance to full precision networks, such as AlexNet, ResNet, GoogLeNet and VGG-Net, than previously available low-precision networks, with 1-bit binary weights and 2-bit quantized activations.
**********

Deep Pyramidal Residual Networks
Dongyoon Han, Jiwhan Kim, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5927-5935
Deep convolutional neural networks (DCNNs) have shown remarkable performance in image classification tasks in recent years. Generally, deep neural network architectures are stacks consisting of a large number of convolutional layers, and they perform downsampling along the spatial dimension via pooling to reduce memory usage. Concurrently, the feature map dimension (i.e., the number of channels) is sharply increased at downsampling locations, which is essential to ensure effective performance because it increases the diversity of high-level attributes. This also applies to residual networks and is very closely related to their performance. In this research, instead of sharply increasing the feature map dimension at units that perform downsampling, we gradually increase the feature map dimension at all units to involve as many locations as possible. This design, which is discussed in depth together with our new insights, has proven to be an effective means of improving generalization ability. Furthermore, we propose a novel residual unit capable of further improving the classification accuracy with our new network architecture. Experiments on benchmark CIFAR-10, CIFAR-100, and ImageNet datasets have shown that our network architecture has superior generalization ability compared to the original residual networks.
**********

RON: Reverse Connection With Objectness Prior Networks for Object Detection
Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, Yurong Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5936-5944
We present RON, an efficient and effective framework for generic object detection. Our motivation is to smartly associate the best of the region-based (e.g., Faster R-CNN) and region-free (e.g., SSD) methodologies. Under fully convolutional architecture, RON mainly focuses on two fundamental problems: (a) multi-scale object localization and (b) negative sample mining. To address (a), we design the reverse connection, which enables the network to detect objects on multi-levels of CNNs. To deal with (b), we propose the objectness prior to significantly reduce the searching space of objects. We optimize the reverse connection, objectness prior and object detector jointly by a multi-task loss function, thus RON can directly predict final detection results from all locations of various feature maps. Extensive experiments on the challenging PASCAL VOC 2007, PASCAL VOC 2012 and MS COCO benchmarks demonstrate the competitive performance of RON. Specifically, with VGG-16 and low resolution 384*384 input size, the network gets 81.3% mAP on PASCAL VOC 2007, 80.7% mAP on PASCAL VOC 2012 datasets. Its superiority increases when datasets become larger and more difficult, as demonstrated by the results on the MS COCO dataset. With 1.5G GPU memory at test phase, the speed of the network is 15 FPS, 3 times faster than the Faster R-CNN counterpart. Code wi

ll be made publicly available.
**********
Weakly-Supervised Visual Grounding of Phrases With Linguistic Structures
Fanyi Xiao, Leonid Sigal, Yong Jae Lee; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2017, pp. 5945-5954
We propose a weakly-supervised approach that takes image-sentence pairs as input
 and learns to visually ground (i.e., localize) arbitrary linguistic phrases, in
 the form of spatial attention masks. Specifically, the model is trained with im
ages and their associated image-level captions, without any explicit region-to-p
hrase correspondence annotations. To this end, we introduce an end-to-end model
which learns visual groundings of phrases with two types of carefully designed l
oss functions. In addition to the standard discriminative loss, which enforces t
hat attended image regions and phrases are consistently encoded, we propose a no
vel structural loss which makes use of the parse tree structures induced by the
sentences. In particular, we ensure complementarity among the attention masks th
at correspond to sibling noun phrases, and compositionality of attention masks a
mong the children and parent phrases, as defined by the sentence parse tree. We
validate the effectiveness of our approach on the Microsoft COCO and Visual Geno
me datasets.
**********
Network Sketching: Exploiting Binary Structure in Deep CNNs
Yiwen Guo, Anbang Yao, Hao Zhao, Yurong Chen; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5955-5963
Convolutional neural networks (CNNs) with deep architectures have substantially
advanced the state-of-the-art in computer vision tasks. However, deep networks a
re typically resource-intensive and thus difficult to be deployed on mobile devi
ces. Recently, CNNs with binary weights have shown compelling efficiency to the
community, whereas the accuracy of such models is usually unsatisfactory in prac
tice. In this paper, we introduce network sketching as a novel technique of purs
uing binary-weight CNNs, targeting at more faithful inference and better trade-o
ff for practical applications. Our basic idea is to exploit binary structure dir
ectly in pre-trained filter banks and to produce binary-weight models via tensor
 expansion. The whole process can be treated as a coarse-to-fine model approxima
tion, akin to the pencil drawing steps of outlining and shading. To further spee
dup the generated models, namely the sketches, we also propose an associative im
plementation of binary tensor convolutions. Experimental results demonstrate tha
t a proper sketch of AlexNet (or ResNet) outperforms the existing binary-weight
models by large margins on the ImageNet large scale classification task, while t
he committed memory for network parameters only exceeds a little.
**********
CASENet: Deep Category-Aware Semantic Edge Detection
Zhiding Yu, Chen Feng, Ming-Yu Liu, Srikumar Ramalingam; Proceedings of the IEEE
 Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5964-59
73
Boundary and edge cues are highly beneficial in improving a wide variety of visi
on tasks such as semantic segmentation, object recognition, stereo, and object p
roposal generation. Recently, the problem of edge detection has been revisited a
nd significant progress has been made with deep learning. While classical edge d
etection is a challenging binary problem in itself, the category-aware semantic
edge detection by nature is an even more challenging multi-label problem. We mod
el the problem such that each edge pixel can be associated with more than one cl
ass as they appear in contours or junctions belonging to two or more semantic cl
asses. To this end, we propose a novel end-to-end deep semantic edge learning ar
chitecture based on ResNet and a new skip-layer architecture where category-wise
 edge activations at the top convolution layer share and are fused with the same
 set of bottom layer features. We then propose a multi-label loss function to su
pervise the fused activations. We show that our proposed architecture benefits t
his problem with better performance, and we outperform the current state-of-the-
art semantic edge detection methods by a large margin on standard data sets such
 as SBD and Cityscapes.

**********
Geometric Loss Functions for Camera Pose Regression With Deep Learning
Alex Kendall, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2017, pp. 5974-5983
Deep learning has shown to be effective for robust and real-time monocular image
 relocalisation. In particular, PoseNet is a deep convolutional neural network w
hich learns to regress the 6-DOF camera pose from a single image. It learns to l
ocalize using high level features and is robust to difficult lighting, motion bl
ur and unknown camera intrinsics, where point based SIFT registration fails. How
ever, it was trained using a naive loss function, with hyper-parameters which re
quire expensive tuning. In this paper, we give the problem a more fundamental th
eoretical treatment. We explore a number of novel loss functions for learning ca
mera pose which are based on geometry and scene reprojection error. Additionally
 we show how to automatically learn an optimal weighting to simultaneously regre
ss position and orientation. By leveraging geometry, we demonstrate that our tec
hnique significantly improves PoseNet's performance across datasets ranging from
 indoor rooms to a small city.
**********
Model-Based Iterative Restoration for Binary Document Image Compression With Dic
tionary Learning
Yandong Guo, Cheng Lu, Jan P. Allebach, Charles A. Bouman; Proceedings of the IE
EE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5984-
5993
The inherent noise in the observed (e.g., scanned) binary document image degrade
s the image quality and harms the compression ratio through breaking the pattern
 repentance and adding entropy to the document images. In this paper, we design
a cost function in Bayesian framework with dictionary learning. Minimizing our c
ost function produces a restored image which has better quality than that of the
 observed noisy image, and a dictionary for representing and encoding the image.
 After the restoration, we use this dictionary (from the same cost function) to
encode the restored image following the symbol-dictionary framework by JBIG2 sta
ndard with the lossless mode. Experimental results with a variety of document im
ages demonstrate that our method improves the image quality compared with the ob
served image, and simultaneously improves the compression ratio. For the test im
ages with synthetic noise, our method reduces the number of flipped pixels by 48
.2% and improves the compression ratio by 36.36% as compared with the best encod
ing methods. For the test images with real noise, our method visually improves t
he image quality, and outperforms the cutting-edge method by 28.27% in terms of
the compression ratio.
**********
Fine-Grained Image Classification via Combining Vision and Language
Xiangteng He, Yuxin Peng; Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition (CVPR), 2017, pp. 5994-6002
Fine-grained image classification is a challenging task due to the large intra-c
lass variance and small inter-class variance, aiming at recognizing hundreds of
sub-categories belonging to the same basic-level category. Most existing fine-gr
ained image classification methods generally learn part detection models to obta
in the semantic parts for better classification accuracy. Despite achieving prom
ising results, these methods mainly have two limitations: (1) not all the parts
which obtained through the part detection models are beneficial and indispensabl
e for classification, and (2) fine-grained image classification requires more de
tailed visual descriptions which could not be provided by the part locations or
attribute annotations. For addressing the above two limitations, this paper prop
oses the two-stream model combing vision and language (CVL) for learning latent
semantic representations. The vision stream learns deep representations from the
 original visual information via deep convolutional neural network. The language
 stream utilizes the natural language descriptions which could point out the dis
criminative parts or characteristics for each image, and provides a flexible and
 compact way of encoding the salient visual aspects for distinguishing sub-categ
ories. Since the two streams are complementary, combing the two streams can furt

her achieves better classification accuracy. Comparing with 12 state-of-the-art methods on the widely used CUB-200-2011 dataset for fine-grained image classification, the experimental results demonstrate our CVL approach achieves the best performance.
**********

A Minimal Solution for Two-View Focal-Length Estimation Using Two Affine Correspondences
Daniel Barath, Tekla Toth, Levente Hajder; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6003-6011
A minimal solution using two affine correspondences is presented to estimate the common focal length and the fundamental matrix between two semi-calibrated cameras - known intrinsic parameters except a common focal length. To the best of our knowledge, this problem is unsolved. The proposed approach extends point correspondence-based techniques with linear constraints derived from local affine transformations. The obtained multivariate polynomial system is efficiently solved by the hidden-variable technique. Observing the geometry of local affinities, we introduce novel conditions eliminating invalid roots. To select the best one out of the remaining candidates, a root selection technique is proposed outperforming the recent ones especially in case of high-level noise. The proposed 2-point algorithm is validated on both synthetic data and 104 publicly available real image pairs. A Matlab implementation of the proposed solution is included in the paper.
**********

Joint Graph Decomposition & Node Labeling: Problem, Algorithms, Applications
Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, Bjoern Andres; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6012-6020
We state a combinatorial optimization problem whose feasible solutions define both a decomposition and a node labeling of a given graph. This problem offers a common mathematical abstraction of seemingly unrelated computer vision tasks, including instance-separating semantic segmentation, articulated human body pose estimation and multiple object tracking. Conceptually, it generalizes the unconstrained integer quadratic program and the minimum cost lifted multicut problem, both of which are NP-hard. In order to find feasible solutions efficiently, we define two local search algorithms that converge monotonously to a local optimum, offering a feasible solution at any time. To demonstrate the effectiveness of these algorithms in tackling computer vision tasks, we apply them to instances of the problem that we construct from published data, using published algorithms. We report state-of-the-art application-specific accuracy in the three above-mentioned applications.
**********

Detangling People: Individuating Multiple Close People and Their Body Parts via Region Assembly
Hao Jiang, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6021-6029
Today's person detection methods work best when people are in common upright poses and appear reasonably well spaced out in the image. However, in many real images, that's not what people do. People often appear quite close to each other, e.g., with limbs linked or heads touching, and their poses are often not pedestrian-like. We propose an approach to detangle people in multi-person images. We formulate the task as a region assembly problem. Starting from a large set of overlapping regions from body part semantic segmentation and generic object proposals, our optimization approach reassembles those pieces together into multiple person instances. Since optimal region assembly is a challenging combinatorial problem, we present a Lagrangian relaxation method to accelerate the lower bound estimation, thereby enabling a fast branch and bound solution for the global optimum. As output, our method produces a pixel-level map indicating both 1) the body part labels (arm, leg, torso, and head), and 2) which parts belong to which individual person. Our results on challenging datasets show our method is robust to

clutter, occlusion, and complex poses. It outperforms a variety of competing methods, including existing detector CRF methods and region CNN approaches. In addition, we demonstrate its impact on a proxemics recognition task, which demands a precise representation of "whose body part is where" in crowded images.
**********

Flight Dynamics-Based Recovery of a UAV Trajectory Using Ground Cameras
Artem Rozantsev, Sudipta N. Sinha, Debadeepta Dey, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6030-6039
We propose a new method to estimate the 6-dof trajectory of a flying object such as a quadrotor UAV within a 3D airspace monitored using multiple fixed ground cameras. It is based on a new structure from motion formulation for the 3D reconstruction of a single moving point with known motion dynamics. Our main contribution is a new bundle adjustment procedure, which in addition to optimizing the camera poses, regularizes the point trajectory using a prior based on motion dynamics (or specifically flight dynamics). Furthermore, we can infer the underlying control input sent to the UAV's autopilot that determined its flight trajectory.
Our method requires neither perfect single-view tracking nor appearance matching across views. For robustness, we allow the tracker to generate multiple detections per frame in each video. The true detections and the data association across videos is estimated using robust multi-view triangulation and subsequently refined in our bundle adjustment formulation. Quantitative evaluation on simulated data and experiments on real videos from indoor and outdoor scenes shows that our technique is superior to existing methods.
**********

SurfNet: Generating 3D Shape Surfaces Using Deep Residual Networks
Ayan Sinha, Asim Unmesh, Qixing Huang, Karthik Ramani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6040-6049
3D shape models are naturally parameterized using vertices and faces, i.e, composed on polygons forming a surface. However, current 3D learning paradigms for predictive and generative tasks using convolutional neural networks focus on a voxelized representation of the object. Lifting convolution operators from the traditional 2D to 3D results in high computational overhead with little additional benefit as most of the geometry information is contained on the surface boundary. Here we study the problem of directly generating the 3D shape surface of rigid and non-rigid shapes using deep convolutional neural networks. We develop a procedure to create consistent `geometry images' representing the 3D shape surface of a category of shapes. We then use this consistent representation for category-specific shape generation from a parametric representation or an image by developing novel extensions of deep residual networks for the task of 3D surface generation. Our experiments indicate that our network learns a meaningful representation of shape surfaces allowing it to interpolate between shape orientations and poses, invent new shape surfaces, reconstruct 3D shape surfaces from previously unseen images, and rectify noisy correspondence between 3D shapes belonging to the same class.
**********

Unite the People: Closing the Loop Between 3D and 2D Human Representations
Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, Peter V. Gehler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6050-6059
3D models provide a common ground for different representations of human bodies. In turn, robust 2D estimation has proven to be a powerful tool to obtain 3D fits "in-the-wild". However, depending on the level of detail, it can be hard to impossible to acquire labeled data for training 2D estimators on large scale. We propose a hybrid approach to this problem: with an extended version of the recently introduced SMPLify method, we obtain high quality 3D body model fits for multiple human pose datasets. Human annotators solely sort good and bad fits. This procedure leads to an initial dataset, UP-3D, with rich annotations. With a comprehensive set of experiments, we show how this data can be used to train discriminative models that produce results with an unprecedented level of detail: our mo

dels predict 31 segments and 91 landmark locations on the body. Using the 91 landmark pose estimator, we present state-of-the art results for 3D human pose and shape estimation using an order of magnitude less training data and without assumptions about gender or pose in the fitting procedure. We show that UP-3D can be enhanced with these improved fits to grow in quantity and quality, which makes the system deployable on large scale. The data, code and models are available for research purposes.
**********

Semantically Consistent Regularization for Zero-Shot Recognition
Pedro Morgado, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6060-6069
The role of semantics in zero-shot learning is considered. The effectiveness of previous approaches is analyzed according to the form of supervision provided. While some learn semantics independently, others only supervise the semantic subspace explained by training classes. Thus, the former is able to constrain the whole space but lacks the ability to model semantic correlations. The latter addresses this issue but leaves part of the semantic space unsupervised. This complementarity is exploited in a new convolutional neural network (CNN) framework, which proposes the use of semantics as constraints for recognition.Although a CNN trained for classification has no transfer ability, this can be encouraged by learning an hidden semantic layer together with a semantic code for classification. Two forms of semantic constraints are then introduced. The first is a loss-based regularizer that introduces a generalization constraint on each semantic predictor. The second is a codeword regularizer that favors semantic-to-class mappings consistent with prior semantic knowledge while allowing these to be learned from data. Significant improvements over the state-of-the-art are achieved on several datasets.
**********

Simultaneous Super-Resolution and Cross-Modality Synthesis of 3D Medical Images Using Weakly-Supervised Joint Convolutional Sparse Coding
Yawen Huang, Ling Shao, Alejandro F. Frangi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6070-6079
Magnetic Resonance Imaging (MRI) offers high-resolution in vivo imaging and rich functional and anatomical multimodality tissue contrast. In practice, however, there are challenges associated with considerations of scanning costs, patient comfort, and scanning time that constrain how much data can be acquired in clinical or research studies. In this paper, we explore the possibility of generating high-resolution and multimodal images from low-resolution single-modality imagery. We propose the weakly-supervised joint convolutional sparse coding to simultaneously solve the problems of super-resolution (SR) and cross-modality image synthesis. The learning process requires only a few registered multimodal image pairs as the training set. Additionally, the quality of the joint dictionary learning can be improved using a larger set of unpaired images. To combine unpaired data from different image resolutions/modalities, a hetero-domain image alignment term is proposed. Local image neighborhoods are naturally preserved by operating on the whole image domain (as opposed to image patches) and using joint convolutional sparse coding. The paired images are enhanced in the joint learning process with unpaired data and an additional maximum mean discrepancy term, which minimizes the dissimilarity between their feature distributions. Experiments show that the proposed method outperforms state-of-the-art techniques on both SR reconstruction and simultaneous SR and cross-modality synthesis.
**********

Viraliency: Pooling Local Virality
Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6080-6088
In our overly-connected world, the automatic recognition of virality -- the quality of an image or video to be rapidly and widely spread -- is of crucial importance, and has recently awaken the interest of the computer vision community Concurrently, recent progress in deep learning architectures showed that global (ave

rage) pooling strategies allow to extract class activation maps, which highlight the part of the image most likely to contain a certain class. We extend this concept by introducing a pooling layer that learns the size of the average support: the learned top-N average (LENA) pooling. We hypothesize that the latent concepts (feature maps) describing virality may require such a rich pooling strategy and perform an extensive evaluation to assess the validity of this hypothesis. Moreover, we also appraise the use of objectness maps at predicting and localizing the virality of an image. Experiments are shown in two publicly available data sets annotated for virality.

**********

Generative Attribute Controller With Conditional Filtered Generative Adversarial Networks

Takuhiro Kaneko, Kaoru Hiramatsu, Kunio Kashino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6089-6098

We present a generative attribute controller (GAC), a novel functionality for generating or editing an image while intuitively controlling large variations of an attribute. This controller is based on a novel generative model called the conditional filtered generative adversarial network (CFGAN), which is an extension of the conventional conditional GAN (CGAN) that incorporates a filtering architecture into the generator input. Unlike the conventional CGAN, which represents an attribute directly using an observable variable (e.g., the binary indicator of attribute presence) so its controllability is restricted to attribute labeling (e.g., restricted to an ON or OFF control), the CFGAN has a filtering architecture that associates an attribute with a multi-dimensional latent variable, enabling latent variations of the attribute to be represented. We also define the filtering architecture and training scheme considering controllability, enabling the variations of the attribute to be intuitively controlled using typical controllers (radio buttons and slide bars). We evaluated our CFGAN on MNIST, CUB, and CelebA datasets and show that it enables large variations of an attribute to be not only represented but also intuitively controlled while retaining identity. We also show that the learned latent space has enough expressive power to conduct attribute transfer and attribute-based image retrieval.

**********

Deep Learning on Lie Groups for Skeleton-Based Action Recognition

Zhiwu Huang, Chengde Wan, Thomas Probst, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6099-6108

In recent years, skeleton-based action recognition has become a popular 3D classification problem. State-of-the-art methods typically first represent each motion sequence as a high-dimensional trajectory on a Lie group with an additional dynamic time warping, and then shallowly learn favorable Lie group features. In this paper we incorporate the Lie group structure into a deep network architecture to learn more appropriate Lie group features for 3D action recognition. Within the network structure, we design rotation mapping layers to transform the input Lie group features into desirable ones, which are aligned better in the temporal domain. To reduce the high feature dimensionality, the architecture is equipped with rotation pooling layers for the elements on the Lie group. Furthermore, we propose a logarithm mapping layer to map the resulting manifold data into a tangent space that facilitates the application of regular output layers for the final classification. Evaluations of the proposed network for standard 3D human action recognition datasets clearly demonstrate its superiority over existing shallow Lie group feature learning methods as well as most conventional deep learning methods.

**********

Dynamic Time-Of-Flight

Michael Schober, Amit Adam, Omer Yair, Shai Mazor, Sebastian Nowozin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6109-6118

Time-of-flight (TOF) depth cameras provide robust depth inference at low power requirements in a wide variety of consumer and industrial applications. These cameras reconstruct a single depth frame from a given set of infrared (IR) frames

captured over a very short exposure period. Operating in this mode the camera e ssentially forgets all information previously captured - and performs depth infe rence from scratch for every frame. We challenge this practice and propose usin g previously captured information when inferring depth. An inherent problem we h ave to address is camera motion over this longer period of collecting observatio ns. We derive a probabilistic framework combining a simple but robust model of c amera and object motion, together with an observation model. This combination al lows us to integrate information over multiple frames while remaining robust to rapid changes. Operating the camera in this manner has implications in terms o f both computational efficiency and how information should be captured. We addre ss these two issues and demonstrate a realtime TOF system with robust temporal i ntegration that improves depth accuracy over strong baseline methods including a daptive spatio-temporal filters.

**********

Can Walking and Measuring Along Chord Bunches Better Describe Leaf Shapes?

Bin Wang, Yongsheng Gao, Changming Sun, Michael Blumenstein, John La Salle; Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2017, pp. 6119-6128

Effectively describing and recognizing leaf shapes under arbitrary deformations, particularly from a large database, remains an unsolved problem. In this resear ch, we attempted a new strategy of describing shape by walking along a bunch of chords that pass through the shape to measure the regions trespassed. A novel ch ord bunch walks (CBW) descriptor is developed through the chord walking that eff ectively integrates the shape image function over the walked chord to reflect th e contour features and the inner properties of the shape. For each contour point , the chord bunch groups multiple pairs of chord walks to build a hierarchical f ramework for a coarse-to-fine description. The proposed CBW descriptor is invari ant to rotation, scaling, translation, and mirror transforms. Instead of using t he expensive optimal correspondence based matching, an improved Hausdorff distan ce encoded correspondence information is proposed for efficient yet effective sh ape matching. In experimental studies, the proposed method obtained substantiall y higher accuracies with low computational cost over the benchmarks, which indic ates the research potential along this direction.

**********

Ubernet: Training a Universal Convolutional Neural Network for Low-, Mid-, and H igh-Level Vision Using Diverse Datasets and Limited Memory

Iasonas Kokkinos; Proceedings of the IEEE Conference on Computer Vision and Patt ern Recognition (CVPR), 2017, pp. 6129-6138

In this work we train in an end-to-end manner a convolutional neural network (C NN) that jointly handles low-, mid-, and high-level vision tasks in a unified a rchitecture. Such a network can act like a `swiss knife' for vision tasks; we call it an "UberNet" to indicate its overarching nature. The main contributio n of this work consists in handling challenges that emerge when scaling up to ma ny tasks. We introduce techniques that facilitate (i) training a deep architect ure while relying on diverse training sets and (ii) training many (potentially u nlimited) tasks with a limited memory budget. This allows us to train in an end -to-end manner a unified CNN architecture that jointly handles (a) boundary dete ction (b) normal estimation (c) saliency estimation (d) semantic segmentation (e ) human part segmentation (f) semantic boundary detection, (g) region proposal g eneration and object detection. We obtain competitive performance while jointly addressing all tasks in 0.7 seconds on a GPU. Our system will be made publicly available.

**********

Parametric T-Spline Face Morphable Model for Detailed Fitting in Shape Subspace

Weilong Peng, Zhiyong Feng, Chao Xu, Yong Su; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6139-6147

Pre-learnt subspace methods, e.g., 3DMMs, are significant exploration for the sy nthesis of 3D faces by assuming that faces are in a linear class. However, the h uman face is in a nonlinear manifold, and a new test are always not in the pre-l earnt subspace accurately because of the disparity brought by ethnicity, age, ge

nder, etc. In the paper, we propose a parametric T-spline morphable model (T-spl
ineMM) for 3D face representation, which has great advantages of fitting data fr
om an unknown source accurately. In the model, we describe a face by C^2 T-splin
e surface, and divide the face surface into several shape units (SUs), according
 to facial action coding system (FACS), on T-mesh instead of on the surface dire
ctly. A fitting algorithm is proposed to optimize coefficients of T-spline contr
ol point components along pre-learnt identity and expression subspaces, as well
as to optimize the details in refinement progress. As any pre-learnt subspace is
 not complete to handle the variety and details of faces and expressions, it cov
ers a limited span of morphing. SUs division and detail refinement make the mode
l fitting the facial muscle deformation in a larger span of morphing subspace. W
e conduct experiments on face scan data, kinect data as well as the space-time d
ata to test the performance of detail fitting, robustness to missing data and no
ise, and to demonstrate the effectiveness of our model. Convincing results are i
llustrated to demonstrate the effectiveness of our model compared with the popul
ar methods.
**********

Convolutional Neural Network Architecture for Geometric Matching
Ignacio Rocco, Relja Arandjelovic, Josef Sivic; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6148-6157
We address the problem of determining correspondences between two images in agre
ement with a geometric model such as an affine or thin-plate spline transformati
on, and estimating its parameters. The contributions of this work are three-fold
. First, we propose a convolutional neural network architecture for geometric ma
tching. The architecture is based on three main components that mimic the standa
rd steps of feature extraction, matching and simultaneous inlier detection and m
odel parameter estimation, while being trainable end-to-end. Second, we demonstr
ate that the network parameters can be trained from synthetically generated imag
ery without the need for manual annotation and that our matching layer significa
ntly increases generalization capabilities to never seen before images. Finally,
 we show that the same model can perform both instance-level and category-level
matching giving state-of-the-art results on the challenging Proposal Flow datase
t.
**********

Deep Representation Learning for Human Motion Prediction and Classification
Judith Butepage, Michael J. Black, Danica Kragic, Hedvig Kjellstrom; Proceedings
 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,
 pp. 6158-6166
Generative models of 3D human motion are often restricted to a small number of a
ctivities and can therefore not generalize well to novel movements or applicatio
ns.  In this work we propose a deep learning framework for  human motion captu
re data that learns a generic representation from a large corpus of motion captu
re data and generalizes well to new, unseen, motions. Using an encoding-decoding
 network that learns to predict future 3D poses from the most recent past, we ex
tract a feature representation of human motion. Most work on deep learning for s
equence prediction focuses on video and speech. Since skeletal data has a differ
ent structure, we present and evaluate different network architectures that make
 different assumptions about time dependencies and limb correlations. To quantif
y the learned features, we use the output of different layers for action classif
ication and visualize the receptive fields of the network units. Our method outp
erforms the recent state of the art in skeletal motion prediction even though th
ese use action specific training data. Our results show that deep feedforward ne
tworks, trained from a generic mocap database, can successfully be used for feat
ure extraction from human motion data and that this representation can be used a
s a foundation for classification and prediction.
**********

Deep Affordance-Grounded Sensorimotor Object Recognition
Spyridon Thermos, Georgios Th. Papadopoulos, Petros Daras, Gerasimos Potamianos;
 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (
CVPR), 2017, pp. 6167-6175

It is well-established by cognitive neuroscience that human perception of objects constitutes a complex process, where object appearance information is combined with evidence about the so-called object "affordances", namely the types of actions that humans typically perform when interacting with them. This fact has recently motivated the "sensorimotor" approach to the challenging task of automatic object recognition, where both information sources are fused to improve robustness. In this work, the aforementioned paradigm is adopted, surpassing current limitations of sensorimotor object recognition research. Specifically, the deep learning paradigm is introduced to the problem for the first time, developing a number of novel neuro-biologically and neuro-physiologically inspired architectures that utilize state-of-the-art neural networks for fusing the available information sources in multiple ways. The proposed methods are evaluated using a large RGB-D corpus, which is specifically collected for the task of sensorimotor object recognition and is made publicly available. Experimental results demonstrate the utility of affordance information to object recognition, achieving an up to 29% relative error reduction by its inclusion.

**********

All You Need Is Beyond a Good Init: Exploring Better Solution for Training Extremely Deep Convolutional Neural Networks With Orthonormality and Modulation

Di Xie, Jiang Xiong, Shiliang Pu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6176-6185

Deep neural network is difficult to train and this predicament becomes worse as the depth increases. The essence of this problem exists in the magnitude of back propagated errors that will result in gradient vanishing or exploding phenomenon. We show that a variant of regularizer which utilizes orthonormality among different filter banks can alleviate this problem. Moreover, we design a backward error modulation mechanism based on the quasi-isometry assumption between two consecutive parametric layers. Equipped with these two ingredients, we propose several novel optimization solutions that can be utilized for training a specific-structured (repetitively triple modules of Conv-BNReLU) extremely deep convolutional neural network (CNN) WITHOUT any shortcuts/ identity mappings from scratch. Experiments show that our proposed solutions can achieve distinct improvements for a 44-layer and a 110-layer plain networks on both the CIFAR-10 and ImageNet datasets. Moreover, we can successfully train plain CNNs to match the performance of the residual counterparts.  Besides, we propose new principles for designing network structure from the insights evoked by orthonormality. Combined with residual structure, we achieve comparative performance on the ImageNet dataset.

**********

Scale-Aware Face Detection

Zekun Hao, Yu Liu, Hongwei Qin, Junjie Yan, Xiu Li, Xiaolin Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6186-6195

Convolutional neural network (CNN) based face detectors are inefficient in handling faces of diverse scales. They rely on either fitting a large single model to faces across a large scale range or multi-scale testing. Both are computationally expensive. We propose Scale-aware Face Detection (SAFD) to handle scale explicitly using CNN, and achieve better performance with less computation cost. Prior to detection, an efficient CNN predicts the scale distribution histogram of the faces. Then the scale histogram guides the zoom-in and zoom-out of the image. Since the faces will be approximately in uniform scale after zoom, they can be detected accurately even with much smaller CNN. Actually, more than 99% of the faces in AFW can be covered with less than two zooms per image. Extensive experiments on FDDB, MALF and AFW show advantages of SAFD.

**********

Intrinsic Grassmann Averages for Online Linear and Robust Subspace Learning

Rudrasis Chakraborty, Soren Hauberg, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6196-6204

Principal Component Analysis (PCA) is a fundamental method for estimating a linear subspace approximation to high-dimensional data. Many algorithms exist in literature to achieve a statistically robust version of PCA called RPCA.  In this p

aper, we present a geometric framework for computing the principal linear subspa ces in both situations that amounts to computing the intrinsic average on the sp ace of all subspaces (the Grassmann manifold). Points on this manifold are defi ned as the subspaces spanned by K-tuples of observations. We show that the intri nsic Grassmann average of these subspaces coincide with the principal components of the observations when they are drawn from a Gaussian distribution. Similar r esults are also shown to hold for the RPCA. Further, we propose an efficient onl ine algorithm to do subspace averaging which is of linear complexity in terms of number of samples and has a linear convergence rate. When the data has outlier s, our proposed online robust subspace averaging algorithm shows significant per formance (accuracy and computation time) gain over a recently published RPCA met hods with publicly accessible code. We have demonstrated competitive performanc e of our proposed online subspace algorithm method on one synthetic and two real data sets. Experimental results depicting stability of our proposed method are also presented. Furthermore, on two real outlier corrupted datasets, we present comparison experiments showing lower reconstruction error using our online RPCA algorithm. In terms of reconstruction error and time required, both our algorit hms outperform the competition.

**********

Object Co-Skeletonization With Co-Segmentation

Koteswar Rao Jerripothula, Jianfei Cai, Jiangbo Lu, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6205-6213

Recent advances in the joint processing of images have certainly shown its advan tages over the individual processing. Different from the existing works geared t owards co-segmentation or co-localization, in this paper, we explore a new joint processing topic: co-skeletonization, which is defined as joint skeleton extrac tion of common objects in a set of semantically similar images. Object skeletoni zation in real world images is a challenging problem, because there is no prior knowledge of the object's shape if we consider only a single image. This motivat es us to resort to the idea of object co-skeletonization hoping that the commonn ess prior existing across the similar images may help, just as it does for other joint processing problems such as co-segmentation. Noting that skeleton can pro vide good scribbles for segmentation, and skeletonization, in turn, needs good s egmentation, we propose a coupled framework for co-skeletonization and co-segmen tation tasks so that they are well informed by each other, and benefit each othe r synergistically. Since it is a new problem, we also construct a benchmark data set for the co-skeletonization task. Extensive experiments demonstrate that prop osed method achieves very competitive results.

**********

Product Split Trees

Artem Babenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6214-6222

In this work, we introduce a new kind of spatial partition trees for efficient n earest-neighbor search. Our approach first identifies a set of useful data split ting directions, and then learns a codebook that can be used to encode such dire ctions. We use the product-quantization idea in order to make the effective code book large, the evaluation of scalar products between the query and the encoded splitting direction very fast, and the encoding itself compact. As a result, the proposed data srtucture (Product Split tree) achieves compact clustering of dat a points, while keeping the traversal very efficient. In the nearest-neighbor s earch experiments on high-dimensional data, product split trees achieved state-o f-the-art performance, demonstrating better speed-accuracy tradeoff than other s patial partition trees.

**********

Pose-Aware Person Recognition

Vijay Kumar, Anoop Namboodiri, Manohar Paluri, C. V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6223-6232

Person recognition methods that use multiple body regions have shown significant

improvements over traditional face-based recognition. One of the primary challenges in full-body person recognition is the extreme variation in pose and view point. In this work, (i) we present an approach that tackles pose variations utilizing multiple models that are trained on specific poses, and combined using pose-aware weights during testing. (ii) For learning a person representation, we propose a network that jointly optimizes a single loss over multiple body regions. (iii) Finally, we introduce new benchmarks to evaluate person recognition in diverse scenarios and show significant improvements over previously proposed approaches on all the benchmarks including the photo album setting of PIPA.
**********

Dynamic FAUST: Registering Human Bodies in Motion
Federica Bogo, Javier Romero, Gerard Pons-Moll, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6233-6242
While the ready availability of 3D scan data has influenced research throughout computer vision, less attention has focused on 4D data; that is 3D scans of moving non-rigid objects, captured over time. To be useful for vision research, such 4D scans need to be registered, or aligned, to a common topology. Consequently, extending mesh registration methods to 4D is important. Unfortunately, no ground-truth datasets are available for quantitative evaluation and comparison of 4D registration methods. To address this we create a novel dataset of high-resolution 4D scans of human subjects in motion, captured at 60 fps. We propose a new mesh registration method that uses both 3D geometry and texture information to register all scans in a sequence to a common reference topology. The approach exploits consistency in texture over both short and long time intervals and deals with temporal offsets between shape and texture capture. We show how using geometry alone results in significant errors in alignment when the motions are fast and non-rigid. We evaluate the accuracy of our registration and provide a dataset of 40,000 raw and aligned meshes. Dynamic FAUST extends the popular FAUST dataset to dynamic 4D data, and is available for research purposes at http://dfaust.is.tue.mpg.de.
**********

CNN-SLAM: Real-Time Dense Monocular SLAM With Learned Depth Prediction
Keisuke Tateno, Federico Tombari, Iro Laina, Nassir Navab; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6243-6252
Given the recent advances in depth prediction from Convolutional Neural Networks (CNNs), this paper investigates how predicted depth maps from a deep neural network can be deployed for the goal of accurate and dense monocular reconstruction. We propose a method where CNN-predicted dense depth maps are naturally fused together with depth measurements obtained from direct monocular SLAM, based on a scheme that privileges depth prediction in image locations where monocular SLAM approaches tend to fail, e.g. along low-textured regions, and vice-versa. We demonstrate the use of depth prediction to estimate the absolute scale of the reconstruction, hence overcoming one of the major limitations of monocular SLAM. Finally, we propose a framework to efficiently fuse semantic labels, obtained from a single frame, with dense SLAM, so to yield semantically coherent scene reconstruction from a single view. Evaluation results on two benchmark datasets show the robustness and accuracy of our approach.
**********

Alternating Direction Graph Matching
D. Khue Le-Huu, Nikos Paragios; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6253-6261
In this paper, we introduce a graph matching method that can account for constraints of arbitrary order, with arbitrary potential functions. Unlike previous decomposition approaches that rely on the graph structures, we introduce a decomposition of the matching constraints. Graph matching is then reformulated as a non-convex non-separable optimization problem that can be split into smaller and much-easier-to-solve subproblems, by means of the alternating direction method of multipliers. The proposed framework is modular, scalable, and can be instantiated

into different variants. Two instantiations are studied exploring pairwise and higher-order constraints. Experimental results on widely adopted benchmarks involving synthetic and real examples demonstrate that the proposed solutions outperform existing pairwise graph matching methods, and competitive with the state of the art in higher-order settings.

**********

DUST: Dual Union of Spatio-Temporal Subspaces for Monocular Multiple Object 3D Reconstruction

Antonio Agudo, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6262-6270

We present an approach to reconstruct the 3D shape of multiple deforming objects from incomplete 2D trajectories acquired by a single camera. Additionally, we simultaneously provide spatial segmentation (i.e., we identify each of the objects in every frame) and temporal clustering (i.e., we split the sequence into primitive actions). This advances existing work, which only tackled the problem for one single object and non-occluded tracks. In order to handle several objects at a time from partial observations, we model point trajectories as a union of spatial and temporal subspaces, and optimize the parameters of both modalities, the non-observed point tracks and the 3D shape via augmented Lagrange multipliers. The algorithm is fully unsupervised and results in a formulation which does not need initialization. We thoroughly validate the method on challenging scenarios with several human subjects performing different activities which involve complex motions and close interaction. We show our approach achieves state-of-the-art 3D reconstruction results, while it also provides spatial and temporal segmentation.

**********

Unsupervised Part Learning for Visual Recognition

Ronan Sicre, Yannis Avrithis, Ewa Kijak, Frederic Jurie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6271-6279

Part-based image classification aims at representing categories by small sets of learned discriminative parts, upon which an image representation is built. Considered as a promising avenue a decade ago, this direction has been neglected since the advent of deep neural networks. In this context, this paper brings two contributions: first, this work proceeds one step further compared to recent part-based models (PBM), focusing on how to learn parts without using any labeled data. Instead of learning a set of parts per class, as generally performed in the PBM literature, the proposed approach both constructs a partition of a given set of images into visually similar groups, and subsequently learns a set of discriminative parts per group in a fully unsupervised fashion. This strategy opens the door to the use of PBM in new applications where labeled data are typically not available, such as instance-based image retrieval. Second, this paper shows that despite the recent success of end-to-end models, explicit part learning can still boost classification performance. We experimentally show that our learned parts can help building efficient image representations, which outperform state-of-the art Deep Convolutional Neural Networks (DCNN) on both classification and retrieval tasks.

**********

Parsing Images of Overlapping Organisms With Deep Singling-Out Networks

Victor Yurchenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6280-6288

This work is motivated by the mostly unsolved task of parsing biological images with multiple overlapping articulated model organisms (such as worms or larvae). We present a general approach that separates the two main challenges associated with such data, individual object shape estimation and object groups disentangling. At the core of the approach is a deep feed-forward singling-out network (SON) that is trained to map each local patch to a vectorial descriptor that is sensitive to the characteristics (e.g. shape) of a central object, while being invariant to the variability of all other surrounding elements. Given a SON, a local image patch can be matched to a gallery of isolated elements using their SON-de

scriptors, thus producing a hypothesis about the shape of the central element in that patch. The image-level optimization based on integer programming can then pick a subset of the hypotheses to explain (parse) the whole image and disentangle groups of organisms. While sharing many similarities with existing "analysis-by-synthesis" approaches, our method avoids the need for stochastic search in the high-dimensional configuration space and numerous rendering operations at test-time. We show that our approach can parse microscopy images of three popular model organisms (the C.Elegans roundworms, the Drosophila larvae, and the E.Coli bacteria) even under significant crowding and overlaps between organisms. We speculate that the overall approach is applicable to a wider class of image parsing problems concerned with crowded articulated objects, for which rendering training images is possible.
**********

Deep Multitask Architecture for Integrated 2D and 3D Human Sensing
Alin-Ionut Popa, Mihai Zanfir, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6289-6298
We propose a deep multitask architecture for fully automatic 2d and 3d human sensing (DMHS), including recognition and reconstruction, in monocular images. The system computes the figure-ground segmentation, semantically identifies the human body parts at pixel level, and estimates the 2d and 3d pose of the person. The model supports the joint training of all components by means of multi-task losses where early processing stages recursively feed into advanced ones for increasingly complex calculations, accuracy and robustness. The design allows us to tie a complete training protocol, by taking advantage of multiple datasets that would otherwise restrictively cover only some of the model components: complex 2d image data with no body part labeling and without associated 3d ground truth, or complex 3d data with limited 2d background variability. In detailed experiments based on several challenging 2d and 3d datasets (LSP, HumanEva, Human3.6M), we evaluate the sub-structures of the model, the effect of various types of training data in the multitask loss, and demonstrate that state-of-the-art results can be achieved at all processing levels. We also show that in the wild our monocular RGB architecture is perceptually competitive to a state-of-the art (commercial) Kinect system based on RGB-D data.
**********

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset
Joao Carreira, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6299-6308
The paucity of videos in current action classification datasets (UCF-101 and HMDB-51) has made it difficult to identify good video architectures, as most methods obtain similar performance on existing small-scale benchmarks. This paper re-evaluates state-of-the-art architectures in light of the new Kinetics Human Action Video dataset. Kinetics has two orders of magnitude more data, with 400 human action classes and over 400 clips per class, and is collected from realistic, challenging YouTube videos. We provide an analysis on how current architectures fare on the task of action classification on this dataset and how much performance improves on the smaller benchmark datasets after pre-training on Kinetics. We also introduce a new Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. We show that, after pre-training on Kinetics, I3D models considerably improve upon the state-of-the-art in action classification, reaching 80.2% on HMDB-51 and 97.9% on UCF-101.
**********

Discriminative Correlation Filter With Channel and Spatial Reliability
Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, Matej Kristan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6309-6318
Short-term tracking is an open and challenging problem for which discriminative correlation filters (DCF) have shown excellent performance. We introduce the c

hannel and spatial reliability concepts to DCF tracking and provide a novel lear ning algorithm for its efficient and seamless integration in the filter update a nd the tracking process. The spatial reliability map adjusts the filter support to the part of the object suitable for tracking. This allows tracking of non-re ctangular objects as well as extending the search region. Channel reliability reflects the quality of the learned filter and it is used as a feature weighting coefficient in localization. Experimentally, with only two simple standard fe atures, HOGs and Colornames, the novel CSR-DCF method -- DCF with Channel and Sp atial Reliability -- achieves state-of-the-art results on VOT 2016, VOT 2015 an d OTB. The CSR-DCF runs in real-time on a CPU.
**********

Light Field Reconstruction Using Deep Convolutional Network on EPI
Gaochang Wu, Mandan Zhao, Liangyong Wang, Qionghai Dai, Tianyou Chai, Yebin Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ( CVPR), 2017, pp. 6319-6327
In this paper, we take advantage of the clear texture structure of the epipolar plane image (EPI) in the light field data and model the problem of light field r econstruction from a sparse set of views as a CNN-based angular detail restorati on on EPI. We indicate that one of the main challenges in sparsely sampled light field reconstruction is the information asymmetry between the spatial and angul ar domain, where the detail portion in the angular domain is damaged by undersam pling. To balance the spatial and angular information, the spatial high frequenc y components of an EPI is removed using EPI blur, before feeding to the network. Finally, a non-blind deblur operation is used to recover the spatial detail sup pressed by the EPI blur. We evaluate our approach on several datasets including synthetic scenes, real-world scenes and challenging microscope light field data. We demonstrate the high performance and robustness of the proposed framework co mpared with the state-of-the-arts algorithms. We also show a further application for depth enhancement by using the reconstructed light field.
**********

Noise Robust Depth From Focus Using a Ring Difference Filter
Jaeheung Surh, Hae-Gon Jeon, Yunwon Park, Sunghoon Im, Hyowon Ha, In So Kweon; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2017, pp. 6328-6337
Depth from focus (DfF) is a method of estimating depth of a scene by using the i nformation acquired through the change of the focus of a camera. Within the fram ework of DfF, the focus measure (FM) forms the foundation on which the accuracy of the output is determined. With the result from the FM, the role of a DfF pipe line is to determine and recalculate unreliable measurements while enhancing tho se that are reliable. In this paper, we propose a new FM that more accurately an d robustly measures focus, which we call the "ring difference filter" (RDF). FMs can usually be categorized as confident local methods or noise robust non-local methods. RDF's unique ring-and-disk structure allows it to have the advantageou s sides of both local and non-local FMs. We then describe an efficient pipeline that utilizes the properties that the RDF brings. Our method is able to reproduc e results that are on par with or even better than those of the state-of-the-art , while spending less time in computation.
**********

Improving RANSAC-Based Segmentation Through CNN Encapsulation
Dustin Morley, Hassan Foroosh; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2017, pp. 6338-6347
In this work, we present a method for improving a random sample consensus (RANSA C) based image segmentation algorithm by encapsulating it within a convolutional neural network (CNN). The improvements are gained by gradient descent training on the set of pre-RANSAC filtering and thresholding operations using a novel RAN SAC-based loss function, which is geared toward optimizing the strength of the c orrect model relative to the most convincing false model. Thus, it can be said t hat our loss function trains the network on metrics that directly dictate the su ccess or failure of the final segmentation rather than metrics that are merely c orrelated to the success or failure. We demonstrate successful application of th

is method to a RANSAC method for identifying the pupil boundary in images from t
he CASIA-IrisV3 iris recognition data set, and we expect that this method could
be successfully applied to any RANSAC-based segmentation algorithm.
**********

Bayesian Supervised Hashing
Zihao Hu, Junxuan Chen, Hongtao Lu, Tongzhen Zhang; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6348-6355
Among learning based hashing methods, supervised hashing seeks compact binary re
presentation of the training data to preserve semantic similarities. Recent year
s have witnessed various problem formulations and optimization methods for super
vised hashing. Most of them optimize a form of loss function with a regulization
 term, which can be viewed as a maximum a posterior (MAP) estimation of the hash
ing codes. However, these approaches are prone to overfitting unless hyperparame
ters are tuned carefully. To address this problem, we present a novel fully Baye
sian treatment for supervised hashing problem, named Bayesian Supervised Hashing
 (BSH), in which hyperparameters are automatically tuned during optimization. Ad
ditionally, by utilizing automatic relevance determination (ARD), we can figure
out relative discriminating ability of different hashing bits and select most in
formative bits among them. Experimental results on three real-world image datase
ts with semantic information show that BSH can achieve superior performance over
 state-of-the-art methods with comparable training time.
**********

Mimicking Very Efficient Network for Object Detection
Quanquan Li, Shengying Jin, Junjie Yan; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2017, pp. 6356-6364
Current CNN based object detectors need initialization from pre-trained ImageNet
 classification models, which are usually time-consuming. In this paper, we pres
ent a fully convolutional feature mimic framework to train very efficient CNN ba
sed detectors, which do not need ImageNet pre-training and achieve competitive p
erformance as the large and slow models. We add supervision from high-level feat
ures of the large networks in training to help the small network better learn ob
ject representation. More specifically, we conduct a mimic method for the featur
es sampled from the entire feature map and use a transform layer to map features
 from the small network onto the same dimension of the large network. In trainin
g the small network, we optimize the similarity between features sampled from th
e same region on the feature maps of both networks. Extensive experiments are co
nducted on pedestrian and common object detection tasks using VGG, Inception and
 ResNet. On both Caltech and Pascal VOC, we show that the modified 2.5x accelera
ted Inception network achieves competitive performance as the full Inception Net
work. Our faster model runs at 80 FPS for a 1000x1500 large input with only a mi
nor degradation of performance on Caltech.
**********

3D Menagerie: Modeling the 3D Shape and Pose of Animals
Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, Michael J. Black; Proceedings of
 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp
. 6365-6373
There has been significant work on learning realistic, articulated, 3D models of
 the human body. In contrast, there are few such models of animals, despite many
 applications. The main challenge is that animals are much less cooperative than
 humans. The best human body models are learned from thousands of 3D scans of pe
ople in specific poses, which is infeasible with live animals. Consequently, we
learn our model from a small set of 3D scans of toy figurines in arbitrary poses
. We employ a novel part-based shape model to compute an initial registration to
 the scans. We then normalize their pose, learn a statistical shape model, and r
efine the registrations and the model together. In this way, we accurately align
 animal scans from different quadruped families with very different shapes and p
oses. With the registration to a common template we learn a shape space represen
ting animals including lions, cats, dogs, horses, cows and hippos. Animal shapes
 can be sampled from the model, posed, animated, and fit to data. We demonstrate
 generalization by fitting it to images of real animals including species not se

en in training.
**********
Training Object Class Detectors With Click Supervision
Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6374-6383
Training object class detectors typically requires a large set of images with objects annotated by bounding boxes. However, manually drawing bounding boxes is very time consuming. In this paper we greatly reduce annotation time by proposing center-click annotations: we ask annotators to click on the center of an imaginary bounding box which tightly encloses the object instance. We then incorporate these clicks into existing Multiple Instance Learning techniques for weakly supervised object localization, to jointly localize object bounding boxes over all training images. Extensive experiments on PASCAL VOC 2007 and MS COCO show that: (1) our scheme delivers high-quality detectors, performing substantially better than those produced by weakly supervised techniques, with a modest extra annotation effort; (2) these detectors in fact perform in a range close to those trained from manually drawn bounding boxes; (3) as the center-click task is very fast, our scheme reduces total annotation time by 9x to 18x.
**********
4D Light Field Superpixel and Segmentation
Hao Zhu, Qi Zhang, Qing Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6384-6392
Superpixel segmentation of 2D image has been widely used in many computer vision tasks. However, limited to the Gaussian imaging principle, there is not a thorough segmentation solution to the ambiguity in defocus and occlusion boundary areas. In this paper, we consider the essential element of image pixel, i.e., rays in the light space and propose light field superpixel (LFSP) segmentation to eliminate the ambiguity. The LFSP is first defined mathematically and then a refocus-invariant metric named LFSP self-similarity is proposed to evaluate the segmentation performance. By building a clique system containing 80 neighbors in light field, a robust refocus-invariant LFSP segmentation algorithm is developed. Experimental results on both synthetic and real light field datasets demonstrate the advantages over the state-of-the-arts in terms of traditional evaluation metrics. Additionally the LFSP self-similarity evaluation under different light field refocus levels shows the refocus-invariance of the proposed algorithm.
**********
Joint Sequence Learning and Cross-Modality Convolution for 3D Biomedical Segmentation
Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, Chung-Yang Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6393-6400
Deep learning models such as convolutional neural network have been widely used in 3D biomedical segmentation and achieve state-of-the-art performance. However, most of them often adapt a single modality or stack multiple modalities as different input channels, which ignores the correlations among them. To leverage the multi-modalities, we propose a deep convolution encoder-decoder structure with fusion layers to incorporate different modalities of MRI data. In addition, we exploit convolutional LSTM (convLSTM) to model a sequence of 2D slices, and jointly learn the multi-modalities and convLSTM in an end-to-end manner. To avoid converging to the certain labels, we adopt a re-weighting scheme and two phase training to handle the label imbalance. Experimental results on BRATS-2015 show that our method outperforms state-of-the-art biomedical segmentation approaches.
**********
Multi-Task Clustering of Human Actions by Sharing Information
Xiaoqiang Yan, Shizhe Hu, Yangdong Ye; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6401-6409
Sharing information between multiple tasks can enhance the accuracy of human action recognition systems. However, using shared information to improve multi-task human action clustering has never been considered before, and cannot be achieve

d using existing clustering methods. In this work, we present a novel and effect ive Multi-Task Information Bottleneck (MTIB) clustering method, which is capable of exploring the shared information between multiple action clustering tasks to improve the performance of individual task. Our motivation is that, different a ction collections always share many similar action patterns, and exploiting the shared information can lead to improved performance. Specifically, MTIB generall y formulates this problem as an information loss minimization function. In this function, the shared information can be quantified by the distributional correla tion of clusters in different tasks, which is based on a high-level common vocab ulary constructed through a novel agglomerative information maximization method. Extensive experiments on two kinds of challenging data sets, including realisti c action data sets (HMDB & UCF50, Olympic & YouTube), and cross-view data sets ( IXMAS, WVU), show that the proposed approach compares favorably to the state-of-the-art methods.

**********

Geodesic Distance Descriptors

Gil Shamai, Ron Kimmel; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2017, pp. 6410-6418

The Gromov-Hausdorff (GH) distance is traditionally used for measuring distances between metric spaces. It was adapted for non-rigid shape comparison and matchi ng of isometric surfaces, and is defined as the minimal distortion of embedding one surface into the other, while the optimal correspondence can be described as the map that minimizes this distortion. Solving such a minimization is a hard c ombinatorial problem that requires precomputation and storing of all pairwise ge odesic distances for the matched surfaces. A popular way for compact representat ion of functions on surfaces is by projecting them into the leading eigenfunctio ns of the Laplace-Beltrami Operator (LBO). When truncated, the basis of the LBO is known to be the optimal for representing functions with bounded gradient in a min-max sense. Methods such as Spectral-GMDS exploit this idea to simplify and efficiently approximate a minimization related to the GH distance by operating i n the truncated spectral domain, and obtain state of the art results for matchin g of nearly isometric shapes. However, when considering only a specific set of f unctions on the surface, such as geodesic distances, an optimized basis could be considered as an even better alternative. Moreover, current simplifications of approximating the GH distance introduce errors due to low rank approximations an d relaxations of the permutation matrices. Here, we define the geodesic distance basis, which is optimal for compact approximation of geodesic distances, in ter ms of Frobenius norm. We use the suggested basis to extract the Geodesic Distanc e Descriptor (GDD), which encodes the geodesic distances information as a linear combination of the basis functions. We then show how these ideas can be used to efficiently and accurately approximate the metric spaces matching problem with almost no loss of information. We incorporate recent methods for efficient appro ximation of the proposed basis and descriptor without actually computing and sto ring all geodesic distances. These observations are used to construct a very sim ple and efficient procedure for shape correspondence. Experimental results show that the GDD improves both accuracy and efficiency of state of the art shape mat ching procedures.

**********

Deeply Aggregated Alternating Minimization for Image Restoration

Youngjung Kim, Hyungjoo Jung, Dongbo Min, Kwanghoon Sohn; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6419-6 427

Regularization-based image restoration has remained an active research topic in image processing and computer vision. It often leverages a guidance signal captu red in different fields as an additional cue. In this work, we present a general framework for image restoration, called deeply aggregated alternating minimizat ion (DeepAM). We propose to train deep neural network to advance two of the step s in the conventional AM algorithm: proximal mapping and b-continuation. Both st eps are learned from a large dataset in an end-to-end manner. The proposed frame work enables the convolutional neural networks (CNNs) to operate as a regularize

r in the AM algorithm. We show that our learned regularizer via deep aggregation outperforms the recent data-driven approaches as well as the nonlocal-based methods. The flexibility and effectiveness of our framework are demonstrated in several restoration tasks, including single image denoising, RGB-NIR restoration, and depth super-resolution.
**********
MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network
Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, Lin Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6428-6436
The inability to interpret the model prediction in semantically and visually meaningful ways is a well-known shortcoming of most existing computer-aided diagnosis methods. In this paper, we propose MDNet to establish a direct multimodal mapping between medical images and diagnostic reports that can read images, generate diagnostic reports, retrieve images by symptom descriptions, and visualize attention, to provide justifications of the network diagnosis process. MDNet includes an image model and a language model. The image model is proposed to enhance multi-scale feature ensembles and utilization efficiency. The language model, integrated with our improved attention mechanism, aims to read and explore discriminative image feature descriptions from reports to learn a direct mapping from sentence words to image pixels. The overall network is trained end-to-end by using our developed optimization strategy. Based on a pathology bladder cancer images and its diagnostic reports (BCIDR) dataset, we conduct sufficient experiments to demonstrate that MDNet outperforms comparative baselines. The proposed image model obtains state-of-the-art performance on two CIFAR datasets as well.
**********
Computational Imaging on the Electric Grid
Mark Sheinin, Yoav Y. Schechner, Kiriakos N. Kutulakos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6437-6446
Night beats with alternating current (AC) illumination. By passively sensing this beat, we reveal new scene information which includes: the type of bulbs in the scene, the phases of the electric grid up to city scale, and the light transport matrix. This information yields unmixing of reflections and semi-reflections, nocturnal high dynamic range, and scene rendering with bulbs not observed during acquisition. The latter is facilitated by a database of bulb response functions for a range of sources, which we collected and provide. To do all this, we built a novel coded-exposure high-dynamic-range imaging technique, specifically designed to operate on the grid's AC lighting.
**********
Lip Reading Sentences in the Wild
Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6447-6456
The goal of this work is to recognise phrases and sentences being spoken by a talking face, with or without the audio. Unlike previous works that have focussed on recognising a limited number of words or phrases, we tackle lip reading as an open-world problem - unconstrained natural language sentences, and in the wild videos. Our key contributions are: (1) a 'Watch, Listen, Attend and Spell' (WLAS) network that learns to transcribe videos of mouth motion to characters; (2) a curriculum learning strategy to accelerate training and to reduce overfitting; (3) a 'Lip Reading Sentences' (LRS) dataset for visual speech recognition, consisting of over 100,000 natural sentences from British television. The WLAS model trained on the LRS dataset surpasses the performance of all previous work on standard lip reading benchmark datasets, often by a significant margin. This lip reading performance beats a professional lip reader on videos from BBC television, and we also demonstrate that if audio is available, then visual information helps to improve speech recognition performance.
**********
ArtTrack: Articulated Multi-Person Tracking in the Wild

Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6457-6465

In this paper we propose an approach for articulated tracking of multiple people in unconstrained videos. Our starting point is a model that resembles existing architectures for single-frame pose estimation but is substantially faster. We achieve this in two ways: (1) by simplifying and sparsifying the body-part relationship graph and leveraging recent methods for faster inference, and (2) by offloading a substantial share of computation onto a feed-forward convolutional architecture that is able to detect and associate body joints of the same person even in clutter. We use this model to generate proposals for body joint locations and formulate articulated tracking as spatio-temporal grouping of such proposals. This allows to jointly solve the association problem for all people in the scene by propagating evidence from strong detections through time and enforcing constraints that each proposal can be assigned to one person only. We report results on a public "MPII Human Pose" benchmark and on a new "MPII Video Pose" dataset of image sequences with multiple people. We demonstrate that our model achieves state-of-the-art results while using only a fraction of time and is able to leverage temporal information to improve state-of-the-art for crowded scenes.

**********

LSTM Self-Supervision for Detailed Behavior Analysis

Biagio Brattoli, Uta Buchler, Anna-Sophia Wahl, Martin E. Schwab, Bjorn Ommer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6466-6475

Behavior analysis provides a crucial non-invasive and easily accessible diagnostic tool for biomedical research. A detailed analysis of posture changes during skilled motor tasks can reveal distinct functional deficits and their restoration during recovery. Our specific scenario is based on a neuroscientific study of rodents recovering from a large sensorimotor cortex stroke and skilled forelimb grasping is being recorded. Given large amounts of unlabeled videos that are recorded during such long-term studies, we seek an approach that captures fine-grained details of posture and its change during rehabilitation without costly manual supervision. Therefore, we utilize self-supervision to automatically learn accurate posture and behavior representations for analyzing motor function. Learning our model depends on the following fundamental elements: (i) limb detection based on a fully convolutional network is ini- tialized solely using motion information, (ii) a novel self- supervised training of LSTMs using only temporal permu- tation yields a detailed representation of behavior, and (iii) back-propagation of this sequence representation also im- proves the description of individual postures. We establish a novel test dataset with expert annotations for evaluation of fine-grained behavior analysis. Moreover, we demonstrate the generality of our approach by successfully applying it to self-supervised learning of human posture on two standard benchmark datasets.

**********

Making 360deg Video Watchable in 2D: Learning Videography for Click Free Viewing

Yu-Chuan Su, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6476-6484

360deg video requires human viewers to actively control "where" to look while watching the video. Although it provides a more immersive experience of the visual content, it also introduces additional burden for viewers; awkward interfaces to navigate the video lead to suboptimal viewing experiences. Virtual cinematography is an appealing direction to remedy these problems, but conventional methods are limited to virtual environments or rely on hand-crafted heuristics. We propose a new algorithm for virtual cinematography that automatically controls a virtual camera within a 360deg video. Compared to the state of the art, our algorithm allows more general camera control, avoids redundant outputs, and extracts its output videos substantially more efficiently. Experimental results on over 7 hours of real "in the wild" video show that our generalized camera control is crucial for viewing 360deg video, while the proposed efficient algorithm is essential for making the generalized control computationally tractable.

**********
Creativity: Generating Diverse Questions Using Variational Autoencoders
Unnat Jain, Ziyu Zhang, Alexander G. Schwing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6485-6494
Generating diverse questions for given images is an important task for computational education, entertainment and AI assistants. Different from many conventional prediction techniques is the need for algorithms to generate a diverse set of plausible questions, which we refer to as "creativity". In this paper we propose a creative algorithm for visual question generation which combines the advantages of variational autoencoders with long short-term memory networks. We demonstrate that our framework is able to generate a large set of varying questions given a single input image.
**********
Tracking by Natural Language Specification
Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, Arnold W.M. Smeulders; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6495-6503
This paper strives to track a target object in a video. Rather than specifying the target in the first frame of a video by a bounding box, we propose to track the object based on a natural language specification of the target, which provides a more natural human-machine interaction as well as a means to improve tracking results. We define three variants of tracking by language specification: one relying on lingual target specification only, one relying on visual target specification based on language, and one leveraging their joint capacity. To show the potential of tracking by natural language specification we extend two popular tracking datasets with lingual descriptions and report experiments. Finally, we also sketch new tracking scenarios in surveillance and other live video streams that become feasible with a lingual specification of the target.
**********
Video Captioning With Transferred Semantic Attributes
Yingwei Pan, Ting Yao, Houqiang Li, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6504-6512
Automatically generating natural language descriptions of videos plays a fundamental challenge for computer vision community. Most recent progress in this problem has been achieved through employing 2-D and/or 3-D Convolutional Neural Networks (CNNs) to encode video content and Recurrent Neural Networks (RNNs) to decode a sentence. In this paper, we present Long Short-Term Memory with Transferred Semantic Attributes (LSTM-TSA)---a novel deep architecture that incorporates the transferred semantic attributes learnt from images and videos into the CNN plus RNN framework, by training them in an end-to-end manner. The design of LSTM-TSA is highly inspired by the facts that 1) semantic attributes play a significant contribution to captioning, and 2) images and videos carry complementary semantics and thus can reinforce each other for captioning. To boost video captioning, we propose a novel transfer unit to model the mutually correlated attributes learnt from images and videos. Extensive experiments are conducted on three public datasets, i.e., MSVD, M-VAD and MPII-MD. Our proposed LSTM-TSA achieves to-date the best published performance in sentence generation on MSVD: 52.8% and 74.0% in terms of BLEU@4 and CIDEr-D. Superior results are also reported on M-VAD and MPII-MD when compared to state-of-the-art methods.
**********
Personalizing Gesture Recognition Using Hierarchical Bayesian Neural Networks
Ajjen Joshi, Soumya Ghosh, Margrit Betke, Stan Sclaroff, Hanspeter Pfister; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6513-6522
Building robust classifiers trained on data susceptible to group or subject-specific variations is a challenging pattern recognition problem. We develop hierarchical Bayesian neural networks to capture subject-specific variations and share statistical strength across subjects. Leveraging recent work on learning Bayesian neural networks, we build fast, scalable algorithms for inferring the posterior distribution over all network weights in the hierarchy. We also develop metho

ds for adapting our model to new subjects when a small number of subject-specifi c personalization data is available. Finally, we investigate active learning alg orithms for interactively labeling personalization data in resource-constrained scenarios. Focusing on the problem of gesture recognition where inter-subject va riations are commonplace, we demonstrate the effectiveness of our proposed techn iques. We test our framework on three widely used gesture recognition datasets, achieving personalization performance competitive with the state-of-the-art.
**********

Flexible Spatio-Temporal Networks for Video Prediction
Chaochao Lu, Michael Hirsch, Bernhard Scholkopf; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6523-6531
We describe a modular framework for video frame prediction. We refer to it as a Flexible Spatio-Temporal Network (FSTN) as it allows the extrapolation of a vide o sequence as well as the estimation of synthetic frames lying in between observ ed frames and thus the generation of slow-motion videos. By devising a customize d objective function comprising decoding, encoding, and adversarial losses, we a re able to mitigate the common problem of blurry predictions, managing to retain high frequency information even for relatively distant future predictions. We p ropose and analyse different training strategies to optimize our model. Extensiv e experiments on several challenging public datasets demonstrate both the versat ility and validity of our model.
**********

Soft-Margin Mixture of Regressions
Dong Huang, Longfei Han, Fernando De la Torre; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6532-6540
Nonlinear regression is a common statistical tool to solve many computer vision problems (e.g., age estimation, pose estimation). Existing approaches to nonlin ear regression fall into two main categories: (1) The universal approach provide s an implicit or explicit homogeneous feature mapping (e.g., kernel ridge regres sion, Gaussian process regression, neural networks). These approaches may fail w hen data is heterogeneous or discontinuous. (2) Divide-and-conquer approaches pa rtition a heterogeneous input feature space and learn multiple local regressors. However, existing divide-and-conquer approaches fail to deal with discontinuiti es between partitions (e.g., Gaussian mixture of regressions) and they cannot gu arantee that the partitioned input space will be homogeneously modeled by local regressors (e.g., ordinal regression). To address these issues, this paper propo ses Soft-Margin Mixture of Regressions (SMMR), a method that directly learns hom ogeneous partitions of the input space and is able to deal with discontinuities. SMMR outperforms the state-of-the-art methods on three popular computer vision tasks: age estimation, crowd counting and viewpoint estimation from images.
**********

Network Dissection: Quantifying Interpretability of Deep Visual Representations
David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6541-6549
We propose a general framework called Network Dissection for quantifying the int erpretability of latent representations of CNNs by evaluating the alignment betw een individual hidden units and a set of semantic concepts. Given any CNN model, the proposed method draws on a data set of concepts to score the semantics of h idden units at each intermediate convolutional layer. The units with semantics a re labeled across a broad range of visual concepts including objects, parts, sce nes, textures, materials, and colors. We use the proposed method to test the hyp othesis that interpretability is an axis-independent property of the representat ion space, then we apply the method to compare the latent representations of var ious networks when trained to solve different classification problems. We furth er analyze the effect of training iterations, compare networks trained with diff erent initializations, and measure the effect of dropout and batch normalization on the interpretability of deep visual representations. We demonstrate that the proposed method can shed light on characteristics of CNN models and training me thods that go beyond measurements of their discriminative power

**********
Straight to Shapes: Real-Time Detection of Encoded Shapes
Saumya Jetley, Michael Sapienza, Stuart Golodetz, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6550-6559
Current object detection approaches predict bounding boxes that provide little instance-specific information beyond location, scale and aspect ratio. In this work, we propose to regress directly to objects' shapes in addition to their bounding boxes and categories. It is crucial to find an appropriate shape representation that is compact and decodable, and in which objects can be compared for higher-order concepts such as view similarity, pose variation and occlusion. To achieve this, we use a denoising convolutional auto-encoder to learn a low-dimensional shape embedding space. We place the decoder network after a fast end-to-end deep convolutional network that is trained to regress directly to the shape vectors provided by the auto-encoder. This yields what to the best of our knowledge is the first real-time shape prediction network, running at 35 FPS on a high-end desktop. With higher-order shape reasoning well-integrated into the network pipeline, the network shows the useful practical quality of generalising to unseen categories that are similar to the ones in the training set, something that most existing approaches fail to handle.
**********
FCSS: Fully Convolutional Self-Similarity for Dense Semantic Correspondence
Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, Kwanghoon Sohn; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6560-6569
We present a descriptor, called fully convolutional self-similarity (FCSS), for dense semantic correspondence. To robustly match points among different instances within the same object class, we formulate FCSS using local self-similarity (LSS) within a fully convolutional network. In contrast to existing CNN-based descriptors, FCSS is inherently insensitive to intra-class appearance variations because of its LSS-based structure, while maintaining the precise localization ability of deep neural networks. The sampling patterns of local structure and the self-similarity measure are jointly learned within the proposed network in an end-to-end and multi-scale manner. As training data for semantic correspondence is rather limited, we propose to leverage object candidate priors provided in existing image datasets and also correspondence consistency between object pairs to enable weakly-supervised learning. Experiments demonstrate that FCSS outperforms conventional handcrafted descriptors and CNN-based descriptors on various benchmarks.
**********
Variational Bayesian Multiple Instance Learning With Gaussian Processes
Manuel Haussmann, Fred A. Hamprecht, Melih Kandemir; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6570-6579
Gaussian Processes (GPs) are effective Bayesian predictors. We here show for the first time that instance labels of a GP classifier can be inferred in the multiple instance learning (MIL) setting using variational Bayes. We achieve this via a new construction of the bag likelihood that assumes a large value if the instance predictions obey the MIL constraints and a small value otherwise. This construction lets us derive the update rules for the variational parameters analytically, assuring both scalable learning and fast convergence. We observe this model to improve the state of the art in instance label prediction from bag-level supervision in the 20 Newsgroups benchmark, as well as in Barrett's cancer tumor localization from histopathology tissue microarray images. Furthermore, we introduce a novel pipeline for weakly supervised object detection naturally complemented with our model, which improves the state of the art on the PASCAL VOC 2007 and 2012 data sets. Last but not least, the performance of our model can be further boosted up using mixed supervision: a combination of weak (bag) and strong (instance) labels.
**********
Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects

Ting Yao, Yingwei Pan, Yehao Li, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6580-6588

Image captioning often requires a large set of training image-sentence pairs. In practice, however, acquiring sufficient training pairs is always expensive, making the recent captioning models limited in their ability to describe objects outside of training corpora (i.e., novel objects). In this paper, we present Long Short-Term Memory with Copying Mechanism (LSTM-C) --- a new architecture that incorporates copying into the Convolutional Neural Networks (CNN) plus Recurrent Neural Networks (RNN) image captioning framework, for describing novel objects in captions. Specifically, freely available object recognition datasets are leveraged to develop classifiers for novel objects. Our LSTM-C then nicely integrates the standard word-by-word sentence generation by a decoder RNN with copying mechanism which may instead select words from novel objects at proper places in the output sentence. Extensive experiments are conducted on both MSCOCO image captioning and ImageNet datasets, demonstrating the ability of our proposed LSTM-C architecture to describe novel objects. Furthermore, superior results are reported when compared to state-of-the-art deep models.

**********

Beyond Instance-Level Image Retrieval: Leveraging Captions to Learn a Global Visual Representation for Semantic Retrieval

Albert Gordo, Diane Larlus; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6589-6598

Querying with an example image is a simple and intuitive interface to retrieve information from a visual database. Most of the research in image retrieval has focused on the task of instance-level image retrieval, where the goal is to retrieve images that contain the same object instance as the query image. In this work we move beyond instance-level retrieval and consider the task of semantic image retrieval in complex scenes, where the goal is to retrieve images that share the same semantics as the query image. We show that, despite its subjective nature, the task of semantically ranking visual scenes is consistently implemented across a pool of human annotators. We also show that a similarity based on human-annotated region-level captions is highly correlated with the human ranking and constitutes a good computable surrogate. Following this observation, we learn a visual embedding of the images where the similarity in the visual space is correlated with their semantic similarity surrogate. We further extend our model to learn a joint embedding of visual and textual cues that allows one to query the database using a text modifier in addition to the query image, adapting the results to the modifier. Finally, our model can ground the ranking decisions by showing regions that contributed the most to the similarity between pairs of images, providing a visual explanation of the similarity.

**********

Fast 3D Reconstruction of Faces With Glasses

Fabio Maninchedda, Martin R. Oswald, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6599-6608

We present a method for the fast 3D face reconstruction of people wearing glasses. Our method explicitly and robustly models the case in which a face to be reconstructed is partially occluded by glasses. We propose a simple and generic model for glasses that copes with a wide variety of different shapes, colors and styles, without the need for any database or learning. Our algorithm is simple, fast and requires only small amounts of both memory and runtime resources, allowing for a fast interactive 3D reconstruction on commodity mobile phones. The thorough evaluation of our approach on synthetic and real data demonstrates superior reconstruction results due to the explicit modeling of glasses.

**********

Non-Local Deep Features for Salient Object Detection

Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, Pierre-Marc Jodoin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6609-6617

Saliency detection aims to highlight the most relevant objects in an image. Methods using conventional models struggle whenever salient objects are pictured on

top of a cluttered background while deep neural nets suffer from excess complex
ity and slow evaluation speeds. In this paper, we propose a simplified convolut
ional neural network which combines local and global information through a mult
i-resolution 4x5 grid structure. Instead of enforcing spacial coherence with a C
RF or superpixels as is usually the case, we implemented a loss function inspire
d by the Mumford-Shah functional which penalizes errors on the boundary. We tra
ined our model on the MSRA-B dataset, and tested it on six different saliency be
nchmark datasets. Results show that our method is on par with the state-of-the-a
rt while reducing computation time by a factor of 18 to 100 times, enabling near
real-time, high performance saliency detection.
**********

Simultaneous Feature Aggregating and Hashing for Large-Scale Image Search
Thanh-Toan Do, Dang-Khoa Le Tan, Trung T. Pham, Ngai-Man Cheung; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.
6618-6627
In most state-of-the-art hashing-based visual search systems, local image descri
ptors of an image are first aggregated as a single feature vector. This feature
vector is then subjected to a hashing function that produces a binary hash code.
In previous work, the aggregating and the hashing processes are designed indepe
ndently. In this paper, we propose a novel framework where feature aggregating a
nd hashing are designed simultaneously and optimized jointly. Specifically, our
joint optimization produces aggregated representations that can be better recons
tructed by some binary codes. This leads to more discriminative binary hash cod
es and improved retrieval accuracy. In addition, we also propose a fast version
of the recently-proposed Binary Autoencoder to be used in our proposed framework
. We perform extensive retrieval experiments on several benchmark datasets with
both SIFT and convolutional features. Our results suggest that the proposed fram
ework achieves significant improvements over the state of the art.
**********

Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucinatio
n and Low-Rank Embedding
Jose Lezama, Qiang Qiu, Guillermo Sapiro; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6628-6637
Surveillance cameras today often capture NIR (near infrared) images in low-light
environments. However, most face datasets accessible for training and verificat
ion are only collected in the VIS (visible light) spectrum. It remains a challen
ging problem to match NIR to VIS face images due to the different light spectrum
. Recently, breakthroughs have been made for VIS face recognition by applying d
eep learning on a huge amount of labeled VIS face samples. The same deep learnin
g approach cannot be simply applied to NIR face recognition for two main reasons
: First, much limited NIR face images are available for training compared to the
VIS spectrum. Second, face galleries to be matched are mostly available only in
the VIS spectrum. In this paper, we propose an approach to extend the deep lea
rning breakthrough for VIS face recognition to the NIR spectrum, without retrain
ing the underlying deep models that see only VIS faces. Our approach consists of
two core components, cross-spectral hallucination and low-rank embedding, to op
timize respectively input and output of a VIS deep model for cross-spectral face
recognition. Cross-spectral hallucination produces VIS faces from NIR images th
rough a deep learning approach. Low-rank embedding restores a low-rank structure
for faces deep features across both NIR and VIS spectrum. We observe that it i
s often equally effective to perform hallucination to input NIR images or low-ra
nk embedding to output deep features for a VIS deep model for cross-spectral rec
ognition. When hallucination and low-rank embedding are deployed together, we ob
serve significant further improvement; we obtain state-of-the-art accuracy on th
e CASIA NIR-VIS v2.0 benchmark, without the need at all to re-train the recognit
ion system.
**********

ECO: Efficient Convolution Operators for Tracking
Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, Michael Felsberg; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,

pp. 6638-6646
In recent years, Discriminative Correlation Filter (DCF) based methods have significantly advanced the state-of-the-art in tracking. However, in the pursuit of ever increasing tracking performance, their characteristic speed and real-time capability have gradually faded. Further, the increasingly complex models, with massive number of trainable parameters, have introduced the risk of severe over-fitting. In this work, we tackle the key causes behind the problems of computational complexity and over-fitting, with the aim of simultaneously improving both speed and performance. We revisit the core DCF formulation and introduce: (i) a factorized convolution operator, which drastically reduces the number of parameters in the model; (ii) a compact generative model of the training sample distribution, that significantly reduces memory and time complexity, while providing better diversity of samples; (iii) a conservative model update strategy with improved robustness and reduced complexity. We perform comprehensive experiments on four benchmarks: VOT2016, UAV123, OTB-2015, and TempleColor. When using expensive deep features, our tracker provides a 20-fold speedup and achieves a 13.0% relative gain in Expected Average Overlap compared to the top ranked method in the VOT2016 challenge. Moreover, our fast variant, using hand-crafted features, operates at 60 Hz on a single CPU, while obtaining 65.0% AUC on OTB-2015.
**********

Semi-Supervised Deep Learning for Monocular Depth Map Prediction
Yevhen Kuznietsov, Jorg Stuckler, Bastian Leibe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6647-6655
Supervised deep learning often suffers from the lack of sufficient training data. Specifically in the context of monocular depth map prediction, it is barely possible to determine dense ground truth depth images in realistic dynamic outdoor environments. When using LiDAR sensors, for instance, noise is present in the distance measurements, the calibration between sensors cannot be perfect, and the measurements are typically much sparser than the camera images. In this paper, we propose a novel approach to depth map prediction from monocular images that learns in a semi-supervised way. While we use sparse ground-truth depth for supervised learning, we also enforce our deep network to produce photoconsistent dense depth maps in a stereo setup using a direct image alignment loss. In experiments we demonstrate superior performance in depth map prediction from single images compared to the state-of-the-art methods.
**********

End-To-End Instance Segmentation With Recurrent Attention
Mengye Ren, Richard S. Zemel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6656-6664
While convolutional neural networks have gained impressive success recently in solving structured prediction problems such as semantic segmentation, it remains a challenge to differentiate individual object instances in the scene. Instance segmentation is very important in a variety of applications, such as autonomous driving, image captioning, and visual question answering. Techniques that combine large graphical models with low-level vision have been proposed to address this problem; however, we propose an end-to-end recurrent neural network (RNN) architecture with an attention mechanism to model a human-like counting process, and produce detailed instance segmentations. The network is jointly trained to sequentially produce regions of interest as well as a dominant object segmentation within each region. The proposed model achieves competitive results on the CVPPP, KITTI, and Cityscapes datasets.
**********

Multigrid Neural Architectures
Tsung-Wei Ke, Michael Maire, Stella X. Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6665-6673
We propose a multigrid extension of convolutional neural networks (CNNs). Rather than manipulating representations living on a single spatial grid, our network layers operate across scale space, on a pyramid of grids. They consume multigrid inputs and produce multigrid outputs; convolutional filters themselves have both within-scale and cross-scale extent. This aspect is distinct from simple mul

tiscale designs, which only process the input at different scales. Viewed in te rms of information flow, a multigrid network passes messages across a spatial py ramid. As a consequence, receptive field size grows exponentially with depth, f acilitating rapid integration of context. Most critically, multigrid structure e nables networks to learn internal attention and dynamic routing mechanisms, and use them to accomplish tasks on which modern CNNs fail. Experiments demonstrate wide-ranging performance advantages of multigrid. On CIFAR and ImageNet classi fication tasks, flipping from a single grid to multigrid within the standard CNN paradigm improves accuracy, while being compute and parameter efficient. Multi grid is independent of other architectural choices; we show synergy in combinati on with residual connections. Multigrid yields dramatic improvement on a synthe tic semantic segmentation dataset. Most strikingly, relatively shallow multigri d networks can learn to directly perform spatial transformation tasks, where, in contrast, current CNNs fail. Together, our results suggest that continuous evo lution of features on a multigrid pyramid is a more powerful alternative to exis ting CNN designs on a flat grid.

**********

Fast Boosting Based Detection Using Scale Invariant Multimodal Multiresolution F iltered Features

Arthur Daniel Costea, Robert Varga, Sergiu Nedevschi; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6674-6683

In this paper we propose a novel boosting-based sliding window solution for obje ct detection which can keep up with the precision of the state-of-the art deep l earning approaches, while being 10 to 100 times faster. The solution takes advan tage of multisensorial perception and exploits information from color, motion an d depth. We introduce multimodal multiresolution filtering of signal intensity, gradient magnitude and orientation channels, in order to capture structure at mu ltiple scales and orientations. To achieve scale invariant classification featur es, we analyze the effect of scale change on features for different filter types and propose a correction scheme. To improve recognition we incorporate 2D and 3 D context by generating spatial, geometric and symmetrical channels. Finally, we evaluate the proposed solution on multiple benchmarks for the detection of pede strians, cars and bicyclists. We achieve competitive results at over 25 frames p er second.

**********

DSAC - Differentiable RANSAC for Camera Localization

Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6684-6692

RANSAC is an important algorithm in robust optimization and a central building b lock for many computer vision applications. In recent years, traditionally hand-crafted pipelines have been replaced by deep learning pipelines, which can be tr ained in an end-to-end fashion. However, RANSAC has so far not been used as part of such deep learning pipelines, because its hypothesis selection procedure is non-differentiable. In this work, we present two different ways to overcome this limitation. The most promising approach is inspired by reinforcement learning, namely to replace the deterministic hypothesis selection by a probabilistic sele ction for which we can derive the expected loss w.r.t. to all learnable paramete rs. We call this approach DSAC, the differentiable counterpart of RANSAC. We app ly DSAC to the problem of camera localization, where deep learning has so far fa iled to improve on traditional approaches. We demonstrate that by directly minim izing the expected loss of the output camera poses, robustly estimated by RANSAC , we achieve an increase in accuracy. In the future, any deep learning pipeline can use DSAC as a robust optimization component.

**********

Group-Wise Point-Set Registration Based on Renyi's Second Order Entropy

Luis G. Sanchez Giraldo, Erion Hasanbelliu, Murali Rao, Jose C. Principe; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6693-6701

In this paper, we describe a set of robust algorithms for group-wise registratio

n using both rigid and non-rigid transformations of multiple unlabelled point-se ts with no bias toward a given set. These methods mitigate the need to establish a correspondence among the point-sets by representing them as probability densi ty functions where the registration is treated as a multiple distribution alignm ent. Holder's and Jensen's inequalities provide a notion of similarity/distance among point-sets and Renyi's second order entropy yields a closed-form solution to the cost function and update equations. We also show that the methods can be improved by normalizing the entropy with a scale factor. These provide simple, f ast and accurate algorithms to compute the spatial transformation function neede d to register multiple point-sets. The algorithms are compared against two well- known methods for group-wise point-set registration. The results show an improve ment in both accuracy and computational complexity.
**********

PoseAgent: Budget-Constrained 6D Object Pose Estimation via Reinforcement Learni ng

Alexander Krull, Eric Brachmann, Sebastian Nowozin, Frank Michel, Jamie Shotton, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Patte rn Recognition (CVPR), 2017, pp. 6702-6710

State-of-the-art computer vision algorithms often achieve efficiency by making d iscrete choices about which hypotheses to explore next. This allows allocation o f computational resources to promising candidates, however, such decisions are n on-differentiable. As a result, these algorithms are hard to train in an end-to -end fashion. In this work we propose to learn an efficient algorithm for the ta sk of 6D object pose estimation. Our system optimizes the parameters of an exist ing state-of-the art pose estimation system using reinforcement learning, where the pose estimation system now becomes the stochastic policy, parametrized by a CNN. Additionally, we present an efficient training algorithm that dramatically reduces computation time. We show empirically that our learned pose estimation p rocedure makes better use of limited resources and improves upon the state-of-th e-art on a challenging dataset. Our approach enables differentiable end-to-end t raining of complex algorithmic pipelines and learns to make optimal use of a giv en computational budget.
**********

MuCaLe-Net: Multi Categorical-Level Networks to Generate More Discriminating Fea tures

Youssef Tamaazousti, Herve Le Borgne, Celine Hudelot; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6711-6720

In a transfer-learning scheme, the intermediate layers of a pre-trained CNN are employed as universal image representation to tackle many visual classification problems. The current trend to generate such representation is to learn a CNN on a large set of images labeled among the most specific categories. Such processe s ignore potential relations between categories, as well as the categorical-leve ls used by humans to classify. In this paper, we propose Multi Categorical-Level Networks (MuCaLe-Net) that include human-categorization knowledge into the CNN learning process. A MuCaLe-Net separates generic categories from each other whil e it independently distinguishes specific ones. It thereby generates different f eatures in the intermediate layers that are complementary when combined together . Advantageously, our method does not require additive data nor annotation to tr ain the network. The extensive experiments over four publicly available benchmar ks of image classification exhibit state-of-the-art performances.
**********

High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis

Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, Hao Li; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6721-6729

Recent advances in deep learning have shown exciting promise in filling large ho les in natural images with semantically plausible and context aware details, imp acting fundamental image manipulation tasks such as object removal. While these learning-based methods are significantly more effective in capturing high-level features than prior techniques, they can only handle very low-resolution inputs

due to memory limitations and difficulty in training. Even for slightly larger images, the inpainted regions would appear blurry and unpleasant boundaries become visible. We propose a multi-scale neural patch synthesis approach based on joint optimization of image content and texture constraints, which not only preserves contextual structures but also produces high-frequency details by matching and adapting patches with the most similar mid-layer feature correlations of a deep classification network. We evaluate our method on the ImageNet and Paris Streetview datasets and achieved state-of-the-art inpainting accuracy. We show our approach produces sharper and more coherent results than prior methods, especially for high-resolution images.
**********

Temporal Attention-Gated Model for Robust Sequence Classification
Wenjie Pei, Tadas Baltrusaitis, David M.J. Tax, Louis-Philippe Morency; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6730-6739
Typical techniques for sequence classification are designed for well-segmented sequences which have been edited to remove noisy or irrelevant parts. Therefore, such methods cannot be easily applied on noisy sequences expected in real-world applications. In this paper, we present the Temporal Attention-Gated Model (TAGM) which integrates ideas from attention models and gated recurrent networks to better deal with noisy or unsegmented sequences. Specifically, we extend the concept of attention model to measure the relevance of each observation (time step) of a sequence. We then use a novel gated recurrent network to learn the hidden representation for the final prediction. An important advantage of our approach is interpretability since the temporal attention weights provide a meaningful value for the salience of each time step in the sequence. We demonstrate the merits of our TAGM approach, both for prediction accuracy and interpretability, on three different tasks: spoken digit recognition, text-based sentiment analysis and visual event recognition.
**********

Multiple-Scattering Microphysics Tomography
Aviad Levis, Yoav Y. Schechner, Anthony B. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6740-6749
Scattering effects in images, including those related to haze, fog and appearance of clouds, are fundamentally dictated by microphysical characteristics of the scatterers. This work defines and derives recovery of these characteristics, in a three-dimensional (3D) heterogeneous medium. Recovery is based on a novel tomography approach. Multi-view (multi-angular) and multi-spectral data are linked to the underlying microphysics using 3D radiative transfer, accounting for multiple-scattering. Despite the nonlinearity of the tomography model, inversion is enabled using a few approximations that we describe. As a case study, we focus on passive remote sensing of the atmosphere, where scatterer retrieval can benefit modeling and forecasting of weather, climate and pollution.
**********

Why You Should Forget Luminance Conversion and Do Something Better
Rang M. H. Nguyen, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6750-6758
One of the most frequently applied low-level operations in computer vision is the conversion of an RGB camera image into its luminance representation. This is also one of the most incorrectly applied operations. Even our most trusted softwares, Matlab and OpenCV, do not perform luminance conversion correctly. In this paper, we examine the main factors that make proper RGB to luminance conversion difficult, in particular: 1) incorrect white-balance, 2) incorrect gamma/tone-curve correction, and 3) incorrect equations. Our analysis shows errors up to 50% for various colors are not uncommon. As a result, we argue that for most computer vision problems there is no need to attempt luminance conversion; instead, there are better alternatives depending on the task.
**********

Deep Quantization: Encoding Convolutional Activations With Deep Generative Model
Zhaofan Qiu, Ting Yao, Tao Mei; Proceedings of the IEEE Conference on Computer V

ision and Pattern Recognition (CVPR), 2017, pp. 6759-6768
Deep convolutional neural networks (CNNs) have proven highly effective for visual recognition, where learning a universal representation from activations of convolutional layer plays a fundamental problem. In this paper, we present Fisher Vector encoding with Variational Auto-Encoder (FV-VAE), a novel deep architecture that quantizes the local activations of convolutional layer in a deep generative model, by training them in an end-to-end manner. To incorporate FV encoding strategy into deep generative models, we introduce Variational Auto-Encoder model, which steers a variational inference and learning in a neural network which can be straightforwardly optimized using standard stochastic gradient method. Different from the FV characterized by conventional generative models (e.g., Gaussian Mixture Model) which parsimoniously fit a discrete mixture model to data distribution, the proposed FV-VAE is more flexible to represent the natural property of data for better generalization. Extensive experiments are conducted on three public datasets, i.e., UCF101, ActivityNet, and CUB-200-2011 in the context of video action recognition and fine-grained image classification, respectively. Superior results are reported when compared to state-of-the-art representations. Most remarkably, our proposed FV-VAE achieves to-date the best published accuracy of 94.2% on UCF101.
**********

Joint Multi-Person Pose Estimation and Semantic Part Segmentation
Fangting Xia, Peng Wang, Xianjie Chen, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6769-6778
Human pose estimation and semantic part segmentation are two complementary tasks in computer vision. In this paper, we propose to solve the two tasks jointly for natural multi-person images, in which the estimated pose provides object-level shape prior to regularize part segments while the part-level segments constrain the variation of pose locations. Specifically, we first train two fully convolutional neural networks (FCNs), namely Pose FCN and Part FCN, to provide initial estimation of pose joint potential and semantic part potential. Then, to refine pose joint location, the two types of potentials are fused with a fully-connected conditional random field (FCRF), where a novel segment-joint smoothness term is used to encourage semantic and spatial consistency between parts and joints. To refine part segments, the refined pose and the original part potential are integrated through a Part FCN, where the skeleton feature from pose serves as additional regularization cues for part segments. Finally, to reduce the complexity of the FCRF, we induce human detection boxes and infer the graph inside each box, making the inference forty times faster.  Since there's no dataset that contains both part segments and pose labels, we extend the PASCAL VOC part dataset with human pose joints and perform extensive experiments to compare our method against several most recent strategies. We show that our algorithm surpasses competing methods by 10.6% in pose estimation with much faster speed and by 1.5% in semantic part segmentation.
**********

DOPE: Distributed Optimization for Pairwise Energies
Jose Dolz, Ismail Ben Ayed, Christian Desrosiers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6779-6788
We formulate an Alternating Direction Method of Multipliers (ADMM) that systematically distributes the computations of any technique for optimizing pairwise functions, including  non-submodular potentials. Such discrete functions are very useful in segmentation and a breadth of other vision problems. Our method decomposes the problem into a large set of small sub-problems, each involving a sub-region of the image domain, which can be solved in parallel. We achieve consistency between the sub-problems through a novel constraint that can be used for a large class of pairwise functions. We give an iterative numerical solution that alternates between solving the sub-problems and updating consistency variables, until convergence. We report comprehensive experiments, which demonstrate the benefit of our general distributed solution in the case of the popular serial algorithm of Boykov and Kolmogorov (BK algorithm) and, also, in the context of non-submodular functions.

**********
Reflectance Adaptive Filtering Improves Intrinsic Image Estimation
Thomas Nestmeyer, Peter V. Gehler; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 6789-6798
Separating an image into reflectance and shading layers poses a challenge for le arning approaches because no large corpus of precise and realistic ground truth decompositions exists. The Intrinsic Images in the Wild (IIW) dataset provides a sparse set of relative human reflectance judgments, which serves as a standard benchmark for intrinsic images. A number of methods use IIW to learn statistical dependencies between the images and their reflectance layer. Although learning plays an important role for high performance, we show that a standard signal pro cessing technique achieves performance on par with current state-of-the-art. We propose a loss function for CNN learning of dense reflectance predictions. Our r esults show a simple pixel-wise decision, without any context or prior knowledge , is sufficient to provide a strong baseline on IIW. This sets a competitive bas eline which only two other approaches surpass. We then develop a joint bilateral filtering method that implements strong prior knowledge about reflectance const ancy. This filtering operation can be applied to any intrinsic image algorithm a nd we improve several previous results achieving a new state-of-the-art on IIW. Our findings suggest that the effect of learning-based approaches may have been over-estimated so far. Explicit prior knowledge is still at least as important t o obtain high performance in intrinsic image decompositions.
**********
DenseReg: Fully Convolutional Dense Shape Regression In-The-Wild
Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefan os Zafeiriou, Iasonas Kokkinos; Proceedings of the IEEE Conference on Computer V ision and Pattern Recognition (CVPR), 2017, pp. 6799-6808
In this paper we propose to learn a mapping from image pixels into a dense templ ate grid through a fully convolutional network. We formulate this task as a re gression problem and train our network by leveraging upon manually annotated fac ial landmarks 'in-the-wild'. We use such landmarks to establish a dense correspo ndence field between a three-dimensional object template and the input image, wh ich then serves as the ground-truth for training our regression system. We show that we can combine ideas from semantic segmentation with regression networks, y ielding a highly-accurate `quantized regression' architecture. Our system, call ed DenseReg, allows us to estimate dense image-to-template correspondences in a fully convolutional manner. As such our network can provide useful correspondenc e information as a stand-alone system, while when used as an initialization for Statistical Deformable Models we obtain landmark localization results that larg ely outperform the current state-of-the-art on the challenging 300W benchmark. We thoroughly evaluate our method on a host of facial analysis tasks, and demons trate its use for other correspondence estimation tasks, such as the human body and the human ear. DenseReg code is made available at http://alpguler.com/DenseR eg.html along with supplementary materials.
**********
Deep Learning Human Mind for Automated Visual Classification
Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim So uly, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pat tern Recognition (CVPR), 2017, pp. 6809-6817
What if we could effectively read the mind and transfer human visual capabilitie s to computer vision methods? In this paper, we aim at addressing this question by developing the first visual object classifier driven by human brain signals. In particular, we employ EEG data evoked by visual object stimuli combined with Recurrent Neural Networks (RNN) to learn a discriminative brain activity manifol d of visual categories in a reading the mind effort. Afterward, we transfer the learned capabilities to machines by training a Convolutional Neural Network (CNN )-based regressor to project images onto the learned manifold, thus allowing mac hines to employ human brain-based features for automated visual classification. We use a 128-channel EEG with active electrodes to record brain activity of seve ral subjects while looking at images of 40 ImageNet object classes. The proposed

RNN-based approach for discriminating object classes using brain signals reaches an average accuracy of about 83%, which greatly outperforms existing methods attempting to learn EEG visual object representations. As for automated object categorization, our human brain-driven approach obtains competitive performance, comparable to those achieved by powerful CNN models and it is also able to generalize over different visual datasets.

**********

Learning Discriminative and Transformation Covariant Local Feature Detectors
Xu Zhang, Felix X. Yu, Svebor Karaman, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6818-6826
Robust covariant local feature detectors are important for detecting local features that are (1) discriminative of the image content and (2) can be repeatably detected at consistent locations when the image undergoes diverse transformations. Such detectors are critical for applications such as image search and scene reconstruction. Many learning-based local feature detectors address one of these two problems while overlooking the other. In this work, we propose a novel learning-based method to simultaneously address both issues. Specifically, we extend the covariant constraint proposed by Lenc and Vedaldi by defining the concepts of "standard patch" and "canonical feature" and leverage these to train a novel robust covariant detector. We show that the introduction of these concepts greatly simplifies the learning stage of the covariant detector, and also makes the detector much more robust. Extensive experiments show that our method outperforms previous hand-crafted and learning-based detectors by large margins in terms of repeatability.

**********

Temporal Action Co-Segmentation in 3D Motion Capture Data and Videos
Konstantinos Papoutsakis, Costas Panagiotakis, Antonis A. Argyros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6827-6836
Given two action sequences, we are interested in spotting/co-segmenting all pairs of sub-sequences that represent the same action. We propose a totally unsupervised solution to this problem. No a-priori model of the actions is assumed to be available. The number of common sub-sequences may be unknown. The sub-sequences can be located anywhere in the original sequences, may differ in duration and the corresponding actions may be performed by a different person, in different style. We treat this type of temporal action co-segmentation as a stochastic optimization problem that is solved by employing Particle Swarm Optimization (PSO). The objective function that is minimized by PSO capitalizes on Dynamic Time Warping (DTW) to compare two action sub-sequences. Due to the generic problem formulation and solution, the proposed method can be applied to motion capture (i.e., 3D skeletal) data or to conventional RGB videos acquired in the wild. We present extensive quantitative experiments on standard data sets as well as on data sets we introduced in this paper. The obtained results demonstrate that the proposed method achieves a remarkable increase in co-segmentation quality compared to all tested state of the art methods.

**********

Learning Diverse Image Colorization
Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, David Forsyth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6837-6845
Colorization is an ambiguous problem, with multiple viable colorizations for a single grey-level image. However, previous methods only produce the single most probable colorization. Our goal is to model the diversity intrinsic to the problem of colorization and produce multiple colorizations that display long-scale spatial co-ordination. We learn a low dimensional embedding of color fields using a variational autoencoder (VAE). We construct loss terms for the VAE decoder that avoid blurry outputs and take into account the uneven distribution of pixel colors. Finally, we build a conditional model for the multi-modal distribution between grey-level image and the color field embeddings. Samples from this conditional model result in diverse colorization. We demonstrate that our method obtains

better diverse colorizations than a standard conditional variational autoencoder (CVAE) model, as well as a recently proposed conditional generative adversarial network (cGAN).
**********

Non-Uniform Subset Selection for Active Learning in Structured Data
Sujoy Paul, Jawadul H. Bappy, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6846-6855
Several works have shown that relationships between data points (i.e., context) in structured data can be exploited to obtain better recognition performance. In this paper, we explore a different, but related, problem: how can these inter-relationships be used to efficiently learn and continuously update a recognition model, with minimal human labeling effort. Towards this goal, we propose an active learning framework to select an optimal subset of data points for manual labeling by exploiting the relationships between them. We construct a graph from the unlabeled data to represent the underlying structure, such that each node represents a data point, and edges represent the inter-relationships between them. Thereafter, considering the flow of beliefs in this graph, we choose those samples for labeling which minimize the joint entropy of the nodes of the graph. This results in significant reduction in manual labeling effort without compromising recognition performance. Our method chooses non-uniform number of samples from each batch of streaming data depending on its information content. Also, the submodular property of our objective function makes it computationally efficient to optimize. The proposed framework is demonstrated in various applications, including document analysis, scene-object recognition, and activity recognition.
**********

VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization
Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, Hongkai Wen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6856-6864
Machine learning techniques, namely convolutional neural networks (CNN) and regression forests, have recently shown great promise in performing 6-DoF localization of monocular images. However, in most cases image-sequences, rather only single images, are readily available. To this extent, none of the proposed learning-based approaches exploit the valuable constraint of temporal smoothness, often leading to situations where the per-frame error is larger than the camera motion. In this paper we propose a recurrent model for performing 6-DoF localization of video-clips. We find that, even by considering only short sequences (20 frames), the pose estimates are smoothed and the localization error can be drastically reduced. Finally, we consider means of obtaining probabilistic pose estimates from our model. We evaluate our method on openly-available real-world autonomous driving and indoor localization datasets.
**********

Hard Mixtures of Experts for Large Scale Weakly Supervised Vision
Sam Gross, Marc'Aurelio Ranzato, Arthur Szlam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6865-6873
Training convolutional networks (CNN's) that fit on a single GPU with minibatch stochastic gradient descent has become effective in practice. However, there is still no effective method for training large networks that do not fit in the memory of a few GPU cards, or for parallelizing CNN training. In this work we show that a simple hard mixture of experts model can be efficiently trained to good effect on large scale hashtag (multilabel) prediction tasks. Mixture of experts models are not new [??], but in the past, researchers have had to devise sophisticated methods to deal with data fragmentation. We show empirically that modern weakly supervised data sets are large enough to support naive partitioning schemes where each data point is assigned to a single expert. Because the experts are independent, training them in parallel is easy, and evaluation is cheap for the size of the model. Furthermore, we show that we can use a single decoding layer for all the experts, allowing a unified feature embedding space. We demonstrate that it is feasible (and in fact relatively painless) to train far larger models than could be practically trained with standard CNN architectures, a

nd that the extra capacity can be well used on current datasets.
**********

Colorization as a Proxy Task for Visual Understanding
Gustav Larsson, Michael Maire, Gregory Shakhnarovich; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6874-6883
We investigate and improve self-supervision as a drop-in replacement for ImageNe t pretraining, focusing on automatic colorization as the proxy task. Self-super vised training has been shown to be more promising for utilizing unlabeled data than other, traditional unsupervised learning methods. We build on this success and evaluate the ability of our self-supervised network in several contexts. On VOC segmentation and classification tasks, we present results that are state-of-the-art among methods not using ImageNet labels for pretraining representations.
Moreover, we present the first in-depth analysis of self-supervision via color ization, concluding that formulation of the loss, training details and network a rchitecture play important roles in its effectiveness. This investigation is fur ther expanded by revisiting the ImageNet pretraining paradigm, asking questions such as: How much training data is needed? How many labels are needed? How much do features change when fine-tuned? We relate these questions back to self-super vision by showing that colorization provides a similarly powerful supervisory si gnal as various flavors of ImageNet pretraining.
**********

A Dataset and Exploration of Models for Understanding Video Data Through Fill-In -The-Blank Question-Answering
Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, Christopher Pal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C VPR), 2017, pp. 6884-6893
While deep convolutional neural networks frequently approach or exceed human-lev el performance in benchmark tasks involving static images, extending this succes s to moving images is not straightforward. Video understanding is of interest fo r many applications, including content recommendation, prediction, summarization , event/object detection, and understanding human visual perception. However, ma ny domains lack sufficient data to explore and perfect video models. In order to address the need for a simple, quantitative benchmark for developing and unders tanding video, we present MovieFIB, a fill-in-the-blank question-answering datas et with over 300,000 examples, based on descriptive video annotations for the vi sually impaired. In addition to presenting statistics and a description of the dataset, we perform a detailed analysis of 5 different models' predictions, and compare these with human performance. We investigate the relative importance of language, static (2D) visual features, and moving (3D) visual features; the effe cts of increasing dataset size, the number of frames sampled; and of vocabulary size. We illustrate that: this task is not solvable by a language model alone; o ur model combining 2D and 3D visual information indeed provides the best result; all models perform significantly worse than human-level. We provide human evalu ation for responses given by different models and find that accuracy on the Movi eFIB evaluation corresponds well with human judgment. We suggest avenues for imp roving video models, and hope that the MovieFIB challenge can be useful for meas uring and encouraging progress in this very interesting field.
**********

Interspecies Knowledge Transfer for Facial Keypoint Detection
Maheen Rashid, Xiuye Gu, Yong Jae Lee; Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition (CVPR), 2017, pp. 6894-6903
We present a method for localizing facial keypoints on animals by transferring k nowledge gained from human faces. Instead of directly finetuning a network train ed to detect keypoints on human faces to animal faces (which is sub-optimal sinc e human and animal faces can look quite different), we propose to first adapt th e animal images to the pre-trained human detection network by correcting for the differences in animal and human face shape. We first find the nearest human nei ghbors for each animal image using an unsupervised shape matching method. We use these matches to train a thin plate spline warping network to warp each animal face to look more human-like. The warping network is then jointly finetuned with

a pre-trained human facial keypoint detection network using an animal dataset. We demonstrate state-of-the-art results on both horse and sheep facial keypoint detection, and significant improvement over simple finetuning, especially when training data is scarce. Additionally, we present a new dataset with 3717 images with horse face and facial keypoint annotations.

**********

Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6904-6913

Problems at the intersection of vision and language are of significant importance both as challenging research questions and for the rich set of applications they enable. However, inherent structure in our world and bias in our language tend to be a simpler signal for learning than visual modalities, resulting in models that ignore visual information, leading to an inflated sense of their capability. We propose to counter these language priors for the task of Visual Question Answering (VQA) and make vision (the V in VQA) matter! Specifically, we balance the popular VQA dataset (Antol et al., ICCV 2015) by collecting complementary images such that every question in our balanced dataset is associated with not just a single image, but rather a pair of similar images that result in two different answers to the question. Our dataset is by construction more balanced than the original VQA dataset and has approximately twice the number of image-question pairs. Our complete balanced dataset is publicly available at http://visualqa.org/ as part of the 2nd iteration of the Visual Question Answering Dataset and Challenge (VQA v2.0). We further benchmark a number of state-of-art VQA models on our balanced dataset. All models perform significantly worse on our balanced dataset, suggesting that these models have indeed learned to exploit language priors. This finding provides the first concrete empirical evidence for what seems to be a qualitative sense among practitioners. Finally, our data collection protocol for identifying complementary images enables us to develop a novel interpretable model, which in addition to providing an answer to the given (image, question) pair, also provides a counter-example based explanation. Specifically, it identifies an image that is similar to the original image, but it believes has a different answer to the same question. This can help in building trust for machines among their users.

**********

Deep Semantic Feature Matching

Nikolai Ufer, Bjorn Ommer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6914-6923

Estimating dense visual correspondences between objects with intra-class variation, deformations and background clutter remains a challenging problem. Thanks to the breakthrough of CNNs there are new powerful features available. Despite their easy accessibility and great success, existing semantic flow methods could not significantly benefit from these without extensive additional training. We introduce a novel method for semantic matching with pre-trained CNN features which is based on convolutional feature pyramids and activation guided feature selection. For the final matching we propose a sparse graph matching framework where each salient feature selects among a small subset of nearest neighbors in the target image. To improve our method in the unconstrained setting without bounding box annotations we introduce novel object proposal based matching constraints. Furthermore, we show that the sparse matching can be transformed into a dense correspondence field. Extensive experimental evaluations on benchmark datasets show that our method significantly outperforms existing semantic matching methods.

**********

Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis

Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6924-6932

The recent work of Gatys et al., who characterized the style of an image by the

statistics of convolutional neural network filters, ignited a renewed interest i
n the texture generation and image stylization problems. While their image gener
ation technique uses a slow optimization process, recently several authors have
proposed to learn generator neural networks that can produce similar outputs in
one quick forward pass. While generator networks are promising, they are still i
nferior in visual quality and diversity compared to generation-by-optimization.
In this work, we advance them in two significant ways. First, we introduce an in
stance normalization module to replace batch normalization with significant impr
ovements to the quality of image stylization. Second, we improve diversity by in
troducing a new learning formulation that encourages generators to sample unbias
edly from the Julesz texture ensemble, which is the equivalence class of all ima
ges characterized by certain filter responses. Together, these two improvements
take feed forward texture synthesis and image stylization much closer to the qua
lity of generation-via-optimization, while retaining the speed advantage.
**********

A Reinforcement Learning Approach to the View Planning Problem
Mustafa Devrim Kaba, Mustafa Gokhan Uzunbas, Ser Nam Lim; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6933-6
941

We present a Reinforcement Learning (RL) solution to the view planning problem (
VPP), which generates a sequence of view points that are capable of sensing all
accessible area of a given object represented as a 3D model. In doing so, the go
al is to minimize the number of view points, making the VPP a class of set cover
ing optimization problem (SCOP). The SCOP is NP-hard, and the inapproximability
results tell us that the greedy algorithm provides the best approximation that r
uns in polynomial time. In order to find a solution that is better than the gree
dy algorithm, (i) we introduce a novel score function by exploiting the geometry
of the 3D model, (ii) we device an intuitive approach to VPP using this score f
unction, and (iii) we cast VPP as a Markovian Decision Process (MDP), and solve
the MDP in RL framework using well-known RL algorithms. In particular, we use SA
RSA, Watkins-Q and TD with function approximation to solve the MDP. We compare t
he results of our method with the baseline greedy algorithm in an extensive set
of test objects, and show that we can outperform the baseline in almost all case
s.
**********

Improving Facial Attribute Prediction Using Semantic Segmentation
Mahdi M. Kalayeh, Boqing Gong, Mubarak Shah; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6942-6950

Attributes are semantically meaningful characteristics whose applicability widel
y crosses category boundaries. They are particularly important in describing and
recognizing concepts where no explicit training example is given, e.g., zero-sh
ot learning. Additionally, since attributes are human describable, they can be u
sed for efficient human-computer interaction. In this paper, we propose to emplo
y semantic segmentation to improve facial attribute prediction. The core idea li
es in the fact that many facial attributes describe local properties. In other w
ords, the probability of an attribute to appear in a face image is far from bein
g uniform in the spatial domain. We build our facial attribute prediction model
jointly with a deep semantic segmentation network. This harnesses the localizati
on cues learned by the semantic segmentation to guide the attention of the attri
bute prediction to the regions where different attributes naturally show up. As
a result of this approach, in addition to recognition, we are able to localize t
he attributes, despite merely having access to image level labels (weak supervis
ion) during training. We evaluate our proposed method on CelebA and LFWA datase
ts and achieve superior results to the prior arts. Furthermore, we show that in
the reverse problem, semantic face parsing improves when facial attributes are a
vailable. That reaffirms the need to jointly model these two interconnected task
s.
**********

Deep Network Flow for Multi-Object Tracking
Samuel Schulter, Paul Vernaza, Wongun Choi, Manmohan Chandraker; Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6951-6960

Data association problems are an important component of many computer vision applications, with multi-object tracking being one of the most prominent examples. A typical approach to data association involves finding a graph matching or network flow that minimizes a sum of pairwise association costs, which are often either hand-crafted or learned as linear functions of fixed features. In this work, we demonstrate that it is possible to learn features for network-flow-based data association via backpropagation, by expressing the optimum of a smoothed network flow problem as a differentiable function of the pairwise association costs. We apply this approach to multi-object tracking with a network flow formulation. Our experiments demonstrate that we are able to successfully learn all cost functions for the association problem in an end-to-end fashion, which outperform hand-crafted costs in all settings. The integration and combination of various sources of inputs becomes easy and the cost functions can be learned entirely from data, alleviating tedious hand-designing of costs.

**********

Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-In-The-Blank Image Captioning

Qing Sun, Stefan Lee, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6961-6969

We develop the first approximate inference algorithm for 1-Best (and M-Best) decoding in bidirectional neural sequence models by extending Beam Search (BS) to reason about both forward and backward time dependencies.  Beam Search (BS) is a widely used approximate inference algorithm for decoding sequences from unidirectional neural sequence models. Interestingly, approximate inference in bidirectional models remains an open problem, despite their significant advantage in modeling information from both the past and future. To enable the use of bidirectional models, we present Bidirectional Beam Search (BiBS), an efficient algorithm for approximate bidirectional inference.  To evaluate our method and as an interesting problem in its own right, we introduce a novel Fill-in-the-Blank Image Captioning task which requires reasoning about both past and future sentence structure to reconstruct sensible image descriptions.  We use this task as well as the Visual Madlibs dataset to demonstrate the effectiveness of our approach, consistently outperforming all baseline methods.

**********

Matting and Depth Recovery of Thin Structures Using a Focal Stack

Chao Liu, Srinivasa G. Narasimhan, Artur W. Dubrawski; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6970-6978

Thin structures such as fence, grass and vessels are common in photography and scientific imaging. They exhibit complex 3D structures with sharp depth variations/discontinuities and mutual occlusions. In this paper, we develop a method to estimate the occlusion matte and depths of thin structures from a focal image stack, which is obtained either by varying the focus/aperture of the lens or computed from a one-shot light field image. We propose an image formation model that explicitly describes the spatially varying optical blur and mutual occlusions for structures located at different depths. Based on the model, we derive an efficient MCMC inference algorithm that enables direct and analytical computations of the iterative update for the model/images without re-rendering images in the sampling process. Then, the depths of the thin structures are recovered using gradient descent with the differential terms computed using the image formation model. We apply the proposed method to scenes at both macro and micro scales. For macro-scale, we evaluate our method on scenes with complex 3D thin structures such as tree branches and grass. For micro-scale, we apply our method to in-vivo microscopic images of micro-vessels with diameters less than 50 um. To our knowledge, the proposed method is the first approach to reconstruct the 3D structures of micro-vessels from non-invasive in-vivo image measurements.

**********

Discovering Causal Signals in Images

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, Leon Bo

ttou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6979-6987
This paper establishes the existence of observable footprints that reveal the "causal dispositions" of the object categories appearing in collections of images. We achieve this goal in two steps. First, we take a learning approach to observational causal discovery, and build a classifier that achieves state-of-the-art performance on finding the causal direction between pairs of random variables, given samples from their joint distribution. Second, we use our causal direction classifier to effectively distinguish between features of objects and features of their contexts in collections of static images. Our experiments demonstrate the existence of a relation between the direction of causality and the difference between objects and their contexts, and by the same token, the existence of observable signals that reveal the causal dispositions of objects.
**********

Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations
Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6988-6997
Recent advances with Convolutional Networks (ConvNets) have shifted the bottleneck for many computer vision tasks to annotated data collection. In this paper, we present a geometry-driven approach to automatically collect annotations for human pose prediction tasks. Starting from a generic ConvNet for 2D human pose, and assuming a multi-view setup, we describe an automatic way to collect accurate 3D human pose annotations. We capitalize on constraints offered by the 3D geometry of the camera setup and the 3D structure of the human body to probabilistically combine per view 2D ConvNet predictions into a globally optimal 3D pose. This 3D pose is used as the basis for harvesting annotations. The benefit of the annotations produced automatically with our approach is demonstrated in two challenging settings: (i) fine-tuning a generic ConvNet-based 2D pose predictor to capture the discriminative aspects of a subject's appearance (i.e.,"personalization"), and (ii) training a ConvNet from scratch for single view 3D human pose prediction without leveraging 3D pose groundtruth. The proposed multi-view pose estimator achieves state-of-the-art results on standard benchmarks, demonstrating the effectiveness of our method in exploiting the available multi-view information.
**********

Shading Annotations in the Wild
Balazs Kovacs, Sean Bell, Noah Snavely, Kavita Bala; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6998-7007
Understanding shading effects in images is critical for a variety of vision and graphics problems, including intrinsic image decomposition, shadow removal, image relighting, and inverse rendering. As is the case with other vision tasks, machine learning is a promising approach to understanding shading - but there is little ground truth shading data available for real-world images. We introduce Shading Annotations in the Wild (SAW), a new large-scale, public dataset of shading annotations in indoor scenes, comprised of multiple forms of shading judgments obtained via crowdsourcing, along with shading annotations automatically generated from RGB-D imagery. We use this data to train a convolutional neural network to predict per-pixel shading information in an image. We demonstrate the value of our data and network in an application to intrinsic images, where we can reduce decomposition artifacts produced by existing algorithms. Our database is available at http://opensurfaces.cs.cornell.edu/saw/.
**********

Self-Critical Sequence Training for Image Captioning
Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, Vaibhava Goel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7008-7024
Recently it has been shown that policy-gradient methods for reinforcement learning can be utilized to train deep end-to-end systems directly on non-differentiable metrics for the task at hand. In this paper we consider the problem of optimizing image captioning systems using reinforcement learning, and show that by car

efully optimizing our systems using the test metrics of the MSCOCO task, signifi
cant gains in performance can be realized. Our systems are built using a new opt
imization approach that we call self-critical sequence training (SCST). SCST is
a form of the popular REINFORCE algorithm that, rather than estimating a "baseli
ne" to normalize the rewards and reduce variance, utilizes the output of its own
 test-time inference algorithm to normalize the rewards it experiences. Using th
is approach, estimating the reward signal (as actor-critic methods must do) and
estimating normalization (as REINFORCE algorithms typically do) is avoided, whil
e at the same time harmonizing the model with respect to its test-time inference
 procedure. Empirically we find that directly optimizing the CIDEr metric with S
CST and greedy decoding at test-time is highly effective.  Our results on the MS
COCO evaluation sever establish a new state-of-the-art on the task, improving th
e best result in terms of CIDEr from 104.9 to 114.7.
**********

Coarse-To-Fine Volumetric Prediction for Single-Image 3D Human Pose
Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, Kostas Daniilidis; Pr
oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP
R), 2017, pp. 7025-7034
This paper addresses the challenge of 3D human pose estimation from a single col
or image. Despite the general success of the end-to-end learning paradigm, top p
erforming approaches employ a two-step solution consisting of a Convolutional Ne
twork (ConvNet) for 2D joint localization and a subsequent optimization step to
recover 3D pose. In this paper, we identify the representation of 3D pose as a c
ritical issue with current ConvNet approaches and make two important contributio
ns towards validating the value of end-to-end learning for this task. First, we
propose a fine discretization of the 3D space around the subject and train a Con
vNet to predict per voxel likelihoods for each joint. This creates a natural rep
resentation for 3D pose and greatly improves performance over the direct regress
ion of joint coordinates. Second, to further improve upon initial estimates, we
employ a coarse-to-fine prediction scheme. This step addresses the large dimensi
onality increase and enables iterative refinement and repeated processing of the
 image features. The proposed approach outperforms all state-of-the-art methods
on standard benchmarks achieving a relative error reduction greater than 30% on
average. Additionally, we investigate using our volumetric representation in a r
elated architecture which is suboptimal compared to our end-to-end approach, but
 is of practical interest, since it enables training when no image with correspo
nding 3D groundtruth is available, and allows us to present compelling results f
or in-the-wild images.
**********

3D Human Pose Estimation = 2D Pose Estimation + Matching
Ching-Hang Chen, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2017, pp. 7035-7043
We explore 3D human pose estimation from a single RGB image. While many approach
es try to directly predict 3D pose from image measurements, we explore a simple
architecture that reasons through intermediate 2D pose predictions. Our approach
 is based on two key observations (1) Deep neural nets have revolutionized 2D po
se estimation, producing accurate 2D predictions even for poses with self-occlus
ions (2) "Big-data"sets of 3D mocap data are now readily available, making it te
mpting to "lift" predicted 2D poses to 3D through simple memorization (e.g., nea
rest neighbors). The resulting architecture is straightforward to implement with
 off-the-shelf 2D pose estimation systems and 3D mocap libraries. Importantly, w
e demonstratethatsuchmethodsoutperformalmostallstate-of-theart 3D pose estimatio
n systems, most of which directly try to regress 3D pose from 2D measurements.
**********

Level Playing Field for Million Scale Face Recognition
Aaron Nech, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2017, pp. 7044-7053
Face recognition has the perception of a solved problem, however when tested at
the million-scale exhibits dramatic variation in accuracies across the different
 algorithms [??]. Are the algorithms very different? Is access to good/big train

ing data their secret weapon? Where should face recognition improve? To address those questions, we created a benchmark, MF2, that requires all algorithms to be trained on same data, and tested at the million scale. MF2 is a public large-scale set with 672K identities and 4.7M photos created with the goal to level playing field for large scale face recognition. We contrast our results with findings from the other two large-scale benchmarks MegaFace Challenge and MS-Celebs-1M where groups were allowed to train on any private/public/big/small set. Some key discoveries: 1) algorithms, trained on MF2, were able to achieve state of the art and comparable results to algorithms trained on massive private sets, 2) some outperformed themselves once trained on MF2, 3) invariance to aging suffers from low accuracies as in MegaFace, identifying the need for larger age variations possibly within identities or adjustment of algorithms in future testing.

**********

Unsupervised Adaptive Re-Identification in Open World Dynamic Camera Networks
Rameswar Panda, Amran Bhuiyan, Vittorio Murino, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7054-7063

Person re-identification is an open and challenging problem in computer vision. Existing approaches have concentrated on either designing the best feature representation or learning optimal matching metrics in a static setting where the number of cameras are fixed in a network. Most approaches have neglected the dynamic and open world nature of the re-identification problem, where a new camera may be temporarily inserted into an existing system to get additional information. To address such a novel and very practical problem, we propose an unsupervised adaptation scheme for re-identification models in a dynamic camera network. First, we formulate a domain perceptive re-identification method based on geodesic flow kernel that can effectively find the best source camera (already installed) to adapt with a newly introduced target camera, without requiring a very expensive training phase. Second, we introduce a transitive inference algorithm for re-identification that can exploit the information from best source camera to improve the accuracy across other camera pairs in a network of multiple cameras. Extensive experiments on four benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-art unsupervised learning based alternatives whilst being extremely efficient to compute.

**********

Deep Feature Interpolation for Image Content Changes
Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, Kilian Weinberger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7064-7073

We propose Deep Feature Interpolation (DFI), a new data- driven baseline for automatic high-resolution image transformation. As the name suggests, DFI relies only on simple linear interpolation of deep convolutional features from pre-trained convnets. We show that despite its simplicity, DFI can perform high-level semantic transformations like "make older/younger", "make bespectacled", "add smile", among others, surprisingly well--sometimes even matching or outperforming the state-of-the-art. This is particularly unexpected as DFI requires no specialized network architecture or even any deep network to be trained for these tasks. DFI therefore can be used as a new baseline to evaluate more complex algorithms and provides a practical answer to the question of which image transformation tasks are still challenging after the advent of deep learning.

**********

3D Bounding Box Estimation Using Deep Learning and Geometry
Arsalan Mousavian, Dragomir Anguelov, John Flynn, Jana Kosecka; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7074-7082

We present a method for 3D object detection and pose estimation from a single image. In contrast to current techniques that only regress the 3D orientation of an object, our method first regresses relatively stable 3D object properties using a deep convolutional neural network and then combines these estimates with geo

metric constraints provided by a 2D object bounding box to produce a complete 3D bounding box. The first network output estimates the 3D object orientation using a novel hybrid discrete-continuous loss, which significantly outperforms the L2 loss. The second output regresses the 3D object dimensions, which have relatively little variance compared to alternatives and can often be predicted for many object types. These estimates, combined with the geometric constraints on translation imposed by the 2D bounding box, enable us to recover a stable and accurate 3D object pose. We evaluate our method on the challenging KITTI object detection benchmark [??] both on the official metric of 3D orientation estimation and also on the accuracy of the obtained 3D bounding boxes. Although conceptually simple, our method outperforms more complex and computationally expensive approaches that leverage semantic segmentation, instance level segmentation and flat ground priors [??] and sub-category detection [??][??]. Our discrete-continuous loss also produces state of the art results for 3D viewpoint estimation on the Pascal 3D+ dataset[??].

**********

Collaborative Summarization of Topic-Related Videos

Rameswar Panda, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7083-7092

Large collections of videos are grouped into clusters by a topic keyword, such as "Eiffel Tower" or "Surfing", with many important visual concepts repeating across them. Such a topically close set of videos have mutual influence on each other, which could be used to summarize one of them by exploiting information from others in the set. We build on this intuition to develop a novel approach to extract a summary that simultaneously captures both important particularities arising in the given video, as well as, generalities identified from the set of videos. The topic-related videos provide visual context to identify the important parts of the video being summarized. We achieve this by developing a collaborative sparse optimization method which can be efficiently solved by a half-quadratic minimization algorithm. Our work builds upon the idea of collaborative techniques from information retrieval and natural language processing, which typically use the attributes of other similar objects to predict the attribute of a given object. Experiments on two challenging and diverse datasets well demonstrate the efficacy of our approach over state-of-the-art methods.

**********

Synthesizing Dynamic Patterns by Spatial-Temporal Generative ConvNet

Jianwen Xie, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7093-7101

Video sequences contain rich dynamic patterns, such as dynamic texture patterns that exhibit stationarity in the temporal domain, and action patterns that are non-stationary in either spatial or temporal domain. We show that a spatial-temporal generative ConvNet can be used to model and synthesize dynamic patterns. The model defines a probability distribution on the video sequence, and the log probability is defined by a spatial-temporal ConvNet that consists of multiple layers of spatial-temporal filters to capture spatial-temporal patterns of different scales. The model can be learned from the training video sequences by an "analysis by synthesis" learning algorithm that iterates the following two steps. Step 1 synthesizes video sequences from the currently learned model. Step 2 then updates the model parameters based on the difference between the synthesized video sequences and the observed training sequences. We show that the learning algorithm can synthesize realistic dynamic patterns.

**********

Comprehension-Guided Referring Expressions

Ruotian Luo, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7102-7111

We consider generation and comprehension of natural language referring expression for objects in an image. Unlike generic "image captioning" which lacks natural standard evaluation criteria, quality of a referring expression may be measured by the receiver's ability to correctly infer which object is being described. Following this intuition, we propose two approaches to utilize models trained for

comprehension task to generate better expressions. First, we use a comprehensio n module trained on human-generated expressions, as a "critic" of referring expr ession generator. The comprehension module serves as a differentiable proxy of h uman evaluation, providing training signal to the generation module. Second, we use the comprehension model in a generate-and-rerank pipeline, which chooses fro m candidate expressions generated by a model according to their performance on t he comprehension task. We show that both approaches lead to improved referring e xpression generation on multiple benchmark datasets.
**********

Zero Shot Learning via Multi-Scale Manifold Regularization
Shay Deutsch, Soheil Kolouri, Kyungnam Kim, Yuri Owechko, Stefano Soatto; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7112-7119
We address zero-shot learning using a new manifold alignment framework based on a localized multi-scale transform on graphs. Our inference approach includes a s moothness criterion for a function mapping nodes on a graph (visual representati on) onto a linear space (semantic representation), which we optimize using multi -scale graph wavelets. The robustness of the ensuing scheme allows us to operate with automatically generated semantic annotations, resulting in an algorithm th at is entirely free of manual supervision, and yet improves the state-of-the-art as measured on benchmark datasets.
**********

LCNN: Lookup-Based Convolutional Neural Network
Hessam Bagherinezhad, Mohammad Rastegari, Ali Farhadi; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7120-7129
Porting state of the art deep learning algorithms to resource constrained comput e platforms (e.g. VR, AR, wearables) is extremely challenging. We propose a fast , compact, and accurate model for convolutional neural networks that enables eff icient learning and inference. We introduce LCNN, a lookup-based convolutional n eural network that encodes convolutions by few lookups to a dictionary that is t rained to cover the space of weights in CNNs. Training LCNN involves jointly lea rning a dictionary and a small set of linear combinations. The size of the dicti onary naturally traces a spectrum of trade-offs between efficiency and accuracy. Our experimental results on ImageNet challenge show that LCNN can offer 3.2x sp eedup while achieving 55.1% top-1 accuracy using AlexNet architecture. Our faste st LCNN offers 37.6x speed up over AlexNet while maintaining 44.3% top-1 accurac y. LCNN not only offers dramatic speed ups at inference, but it also enables eff icient training. In this paper, we show the benefits of LCNN in few-shot learnin g and few-iteration learning, two crucial aspects of on-device training of deep learning models.
**********

Deep Unsupervised Similarity Learning Using Partially Ordered Sets
Miguel A. Bautista, Artsiom Sanakoyeu, Bjorn Ommer; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7130-7139
Unsupervised learning of visual similarities is of paramount importance to compu ter vision, particularly due to lacking training data for fine-grained similarit ies. Deep learning of similarities is often based on relationships between pairs or triplets of samples. Many of these relations are unreliable and mutually con tradicting, implying inconsistencies when trained without supervision informatio n that relates different tuples or triplets to each other. To overcome this prob lem, we use local estimates of reliable (dis-)similarities to initially group sa mples into compact surrogate classes and use local partial orders of samples to classes to link classes to each other. Similarity learning is then formulated as a partial ordering task with soft correspondences of all samples to classes. Ad opting a strategy of self-supervision, a CNN is trained to optimally represent s amples in a mutually consistent manner while updating the classes. The similarit y learning and grouping procedure are integrated in a single model and optimized jointly. The proposed unsupervised approach shows competitive performance on de tailed pose estimation and object classification.
**********

Zero-Shot Classification With Discriminative Semantic Representation Learning
Meng Ye, Yuhong Guo; Proceedings of the IEEE Conference on Computer Vision and P attern Recognition (CVPR), 2017, pp. 7140-7148
Zero-shot learning, a special case of unsupervised domain adaptation where the s ource and target domains have disjoint label spaces, has become increasingly pop ular in the computer vision community. In this paper, we propose a novel zero-sh ot learning method based on discriminative sparse non-negative matrix factorizat ion. The proposed approach aims to identify a set of common high-level semantic components across the two domains via non-negative sparse matrix factorization, while enforcing the representation vectors of the images in this common componen t-based space to be discriminatively aligned with the attribute-based label repr esentation vectors. To fully exploit the aligned semantic information contained in the learned representation vectors of the instances, we develop a label propa gation based testing procedure to classify the unlabeled instances from the unse en classes in the target domain. We conduct experiments on four standard zero-sh ot learning image datasets, by comparing the proposed approach to the state-of-t he-art zero-shot learning methods. The empirical results demonstrate the efficac y of the proposed approach.
**********

Learning Detection With Diverse Proposals
Samaneh Azadi, Jiashi Feng, Trevor Darrell; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7149-7157
To predict a set of diverse and informative proposals with enriched representati ons, this paper introduces a differentiable Determinantal Point Process (DPP) la yer that is able to augment the object detection architectures. Most modern obje ct detection architectures, such as Faster R-CNN, learn to localize objects by m inimizing deviations from the ground truth, but ignore correlation between multi ple proposals and object categories. Non-Maximum Suppression (NMS) as a widely u sed proposal pruning scheme ignores label- and instance-level relations between object candidates resulting in multi-labeled detections. In the multi-class case , NMS selects boxes with the largest prediction scores ignoring the semantic rel ation between categories of potential election. In contrast, our trainable DPP l ayer, allowing for Learning Detection with Diverse Proposals (LDDP), considers b oth label-level contextual information and spatial layout relationships between proposals without increasing the number of parameters of the network, and thus i mproves location and category specifications of final detected bounding boxes su bstantially during both training and inference schemes. Furthermore, we show tha t LDDP keeps it superiority over Faster R-CNN even if the number of proposals ge nerated by LDPP is only  30% as many as those for Faster R-CNN.
**********

Learning Random-Walk Label Propagation for Weakly-Supervised Semantic Segmentati on
Paul Vernaza, Manmohan Chandraker; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2017, pp. 7158-7166
Large-scale training for semantic segmentation is challenging due to the expense  of obtaining training data for this task relative to other vision tasks. We pro pose a novel training approach to address this difficulty. Given cheaply-obtaine d sparse image labelings, we propagate the sparse labels to produce guessed dens e labelings. A standard CNN-based segmentation network is trained to mimic these  labelings. The label-propagation process is defined via random-walk hitting pro babilities, which leads to a differentiable parameterization with uncertainty es timates that are incorporated into our loss. We show that by learning the label- propagator jointly with the segmentation predictor, we are able to effectively l earn semantic edges given no direct edge supervision. Experiments also show that  training a segmentation network in this way outperforms the naive approach.
**********

Adversarial Discriminative Domain Adaptation
Eric Tzeng, Judy Hoffman, Kate Saenko, Trevor Darrell; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7167-7176
Adversarial learning methods are a promising approach to training robust deep ne

tworks, and can generate complex samples across diverse domains. They can also i
mprove recognition despite the presence of domain shift or dataset bias: recent
adversarial approaches to unsupervised domain adaptation reduce the difference b
etween the training and test domain distributions and thus improve generalizatio
n performance. However, while generative adversarial networks (GANs) show compel
ling visualizations, they are not optimal on discriminative tasks and can be lim
ited to smaller shifts. On the other hand, discriminative approaches can handle
larger domain shifts, but impose tied weights on the model and do not exploit a
GAN-based loss. In this work, we first outline a novel generalized framework for
adversarial adaptation, which subsumes recent state-of-the-art approaches as sp
ecial cases, and use this generalized view to better relate prior approaches. We
then propose a previously unexplored instance of our general framework which co
mbines discriminative modeling, untied weight sharing, and a GAN loss, which we
call Adversarial Discriminative Domain Adaptation (ADDA). We show that ADDA is m
ore effective yet considerably simpler than competing domain-adversarial methods
, and demonstrate the promise of our approach by exceeding state-of-the-art unsu
pervised adaptation results on standard domain adaptation tasks as well as a dif
ficult cross-modality object classification task.
**********

An Efficient Background Term for 3D Reconstruction and Tracking With Smooth Surf
ace Models

Mariano Jaimez, Thomas J. Cashman, Andrew Fitzgibbon, Javier Gonzalez-Jimenez, D
aniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern
Recognition (CVPR), 2017, pp. 7177-7185

We present a novel strategy to shrink and constrain a 3D model, represented as a
smooth spline-like surface, within the visual hull of an object observed from o
ne or multiple views. This new 'background' or 'silhouette' term combines the ef
ficiency of previous approaches based on an image-plane distance transform with
the accuracy of formulations based on raycasting or ray potentials. The overall
formulation is solved by alternating an inner nonlinear minization (raycasting)
with a joint optimization of the surface geometry, the camera poses and the data
correspondences. Experiments on 3D reconstruction and object tracking show that
the new formulation corrects several deficiencies of existing approaches, for i
nstance when modelling non-convex shapes. Moreover, our proposal is more robust
against defects in the object segmentation and inherently handles the presence o
f uncertainty in the measurements (e.g. null depth values in images provided by
RGB-D cameras).
**********

The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic
Book Narratives

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, H
al Daume III, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2017, pp. 7186-7195

Visual narrative is often a combination of explicit information and judicious om
issions, relying on the viewer to supply missing details. In comics, most moveme
nts in time and space are hidden in the "gutters" between panels. To follow the
story, readers logically connect panels together by inferring unseen actions thr
ough a process called "closure". While computers can now describe the content of
natural images, in this paper we examine whether they can understand the closur
e-driven narratives conveyed by stylized artwork and dialogue in comic book pane
ls. We collect a dataset, COMICS, that consists of over 1.2 million panels (120
GB) paired with automatic textbox transcriptions. An in-depth analysis of COMICS
demonstrates that neither text nor image alone can tell a comic book story, so
a computer must understand both modalities to keep up with the plot. We introduc
e three cloze-style tasks that ask models to predict narrative and character-cen
tric aspects of a panel given n preceding panels as context. Various deep neural
architectures underperform human baselines on these tasks, suggesting that COMI
CS contains fundamental challenges for both vision and language.
**********

Commonly Uncommon: Semantic Sparsity in Situation Recognition

Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7196-7205

Semantic sparsity is a common challenge in structured visual classification problems; when the output space is complex, the vast majority of the possible predictions are rarely, if ever, seen in the training set. This paper studies semantic sparsity in situation recognition, the task of producing structured summaries of what is happening in images, including activities, objects and the roles objects play within the activity. For this problem, we find empirically that most substructures required for prediction are rare, and current state-of-the-art model performance dramatically decreases if even one such rare substructure exists in the target output.We avoid many such errors by (1) introducing a novel tensor composition function that learns to share examples across substructures more effectively and (2) se- mantically augmenting our training data with automatically gathered examples of rarely observed outputs using web data. When integrated within a complete CRF-based structured prediction model, the tensor-based approach outperforms existing state of the art by a relative improvement of 2.11% and 4.40% on top-5 verb and noun-role accuracy, respectively. Adding 5 million images with our semantic aug- mentation techniques gives further relative improvements of 6.23% and 9.57% on top-5 verb and noun-role accuracy.

**********

Top-Down Visual Saliency Guided by Captions

Vasili Ramanishka, Abir Das, Jianming Zhang, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7206-7215

Neural image/video captioning models can generate accurate descriptions, but their internal process of mapping regions to words is a black box and therefore difficult to explain. Top-down neural saliency methods can find important regions given a high-level semantic task such as object classification, but cannot use a natural language sentence as the top-down input for the task. In this paper, we propose Caption-Guided Visual Saliency to expose the region-to-word mapping in modern encoder-decoder networks and demonstrate that it is learned implicitly from caption training data, without any pixel-level annotations. Our approach can produce spatial or spatiotemporal heatmaps for both predicted captions, and for arbitrary query sentences. It recovers saliency without the overhead of introducing explicit attention layers, and can be used to analyze a variety of existing model architectures and improve their design. Evaluation on large-scale video and image datasets demonstrates that our approach achieves comparable captioning performance with existing methods while providing more accurate saliency heatmaps. Our code is available at visionlearninggroup.github.io/caption-guided-saliency/

**********

The Geometry of First-Returning Photons for Non-Line-Of-Sight Imaging

Chia-Yin Tsai, Kiriakos N. Kutulakos, Srinivasa G. Narasimhan, Aswin C. Sankaranarayanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7216-7224

Non-line-of-sight (NLOS) imaging utilizes the full 5D light transient measurements to reconstruct scenes beyond the camera's field of view. Mathematically, this requires solving an elliptical tomography problem that unmixes the shape and albedo from spatially-multiplexed measurements of the NLOS scene. In this paper, we propose a new approach for NLOS imaging by studying the properties of first-returning photons from three-bounce light paths. We show that the times of flight of first-returning photons are dependent only on the geometry of the NLOS scene and each observation is almost always generated from a single NLOS scene point. Exploiting these properties, we derive a space carving algorithm for NLOS scenes. In addition, by assuming local planarity, we derive an algorithm to localize NLOS scene points in 3D and estimate their surface normals. Our methods do not require either the full transient measurements or solving the hard elliptical tomography problem. We demonstrate the effectiveness of our methods through simulations as well as real data captured from a SPAD sensor.

**********

An Efficient Algebraic Solution to the Perspective-Three-Point Problem
Tong Ke, Stergios I. Roumeliotis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7225-7233
In this work, we present an algebraic solution to the classical perspective-3-point (P3P) problem for determining the position and attitude of a camera from observations of three known reference points. In contrast to previous approaches, we first directly determine the camera's attitude by employing the corresponding geometric constraints to formulate a system of trigonometric equations. This is then efficiently solved, following an algebraic approach, to determine the unknown rotation matrix and subsequently the camera's position. As compared to recent alternatives, our method avoids computing unnecessary (and potentially numerically unstable) intermediate results, and thus achieves higher numerical accuracy and robustness at a lower computational cost. These benefits are validated through extensive Monte-Carlo simulations for both nominal and close-to-singular geometric configurations.
**********

WSISA: Making Survival Prediction From Whole Slide Histopathological Images
Xinliang Zhu, Jiawen Yao, Feiyun Zhu, Junzhou Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7234-7242
Image-based precision medicine techniques can be used to better treat cancer patients. However, the gigapixel resolution of Whole Slide Histopathological Images (WSIs) makes traditional survival models computationally impossible. These models usually adopt manually labeled discriminative patches from region of interests (ROIs) and are unable to directly learn discriminative patches from WSIs. We argue that only a small set of patches cannot fully represent the patients' survival status due to the heterogeneity of tumor. Another challenge is that survival prediction usually comes with insufficient training patient samples. In this paper, we propose an effective Whole Slide Histopathological Images Survival Analysis framework (WSISA) to overcome above challenges. To exploit survival-discriminative patterns from WSIs, we first extract hundreds of patches from each WSI by adaptive sampling and then group these images into different clusters. Then we propose to train an aggregation model to make patient-level predictions based on cluster-level Deep Convolutional Survival (DeepConvSurv) prediction results. Different from existing state-of-the-arts image-based survival models which extract features using some patches from small regions of WSIs, the proposed framework can efficiently exploit and utilize all discriminative patterns in WSIs to predict patients' survival status. To the best of our knowledge, this has not been shown before. We apply our method to the survival predictions of glioma and non-small-cell lung cancer using three datasets. Results demonstrate the proposed framework can significantly improve the prediction performance compared with the existing state-of-the-arts survival methods.
**********

A Low Power, Fully Event-Based Gesture Recognition System
Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, Dharmendra Modha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7243-7252
We present the first gesture recognition system implemented end-to-end on event-based hardware, using a TrueNorth neurosynaptic processor to recognize hand gestures in real-time at low power from events streamed live by a Dynamic Vision Sensor (DVS). The biologically inspired DVS transmits data only when a pixel detects a change, unlike traditional frame-based cameras which sample every pixel at a fixed frame rate. This sparse, asynchronous data representation lets event-based cameras operate at much lower power than frame-based cameras. However, much of the energy efficiency is lost if, as in previous work, the event stream is interpreted by conventional synchronous processors. Here, for the first time, we process a live DVS event stream using TrueNorth, a natively event-based processor with 1 million spiking neurons. Configured here as a convolutional neural network (CNN), the TrueNorth chip identifies the onset of a gesture with a latency of 1

05 ms while consuming less than 200 mW. The CNN achieves 96.5% out-of-sample acc uracy on a newly collected DVS dataset (DvsGesture) comprising 11 hand gesture c ategories from 29 subjects under 3 illumination conditions.
**********

Modeling Sub-Event Dynamics in First-Person Action Recognition
Hasan F. M. Zaki, Faisal Shafait, Ajmal Mian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7253-7262
First-person videos have unique characteristics such as heavy egocentric motion, strong preceding events, salient transitional activities and post-event impacts . Action recognition methods designed for third person videos may not optimally represent actions captured by first-person videos. We propose a method to repres ent the high level dynamics of sub-events in first-person videos by dynamically pooling features of sub-intervals of time series using a temporal feature poolin g function. The sub-event dynamics are then temporally aligned to make a new ser ies. To keep track of how the sub-event dynamics evolve over time, we recursivel y employ the Fast Fourier Transform on a pyramidal temporal structure. The Fouri er coefficients of the segment define the overall video representation. We perfo rm experiments on two existing benchmark first-person video datasets which have been captured in a controlled environment. Addressing this gap, we introduce a n ew dataset collected from YouTube which has a larger number of classes and a gre ater diversity of capture conditions thereby more closely depicting real-world c hallenges in first-person video analysis. We compare our method to state-of-the- art first person and generic video recognition algorithms. Our method consistent ly outperforms the nearest competitors by 10.3%, 3.3% and 11.7% respectively on the three datasets.
**********

YOLO9000: Better, Faster, Stronger
Joseph Redmon, Ali Farhadi; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2017, pp. 7263-7271
We introduce YOLO9000, a state-of-the-art, real-time object detection system tha t can detect over 9000 object categories. First we propose various improvements to the YOLO detection method, both novel and drawn from prior work. The improved model, YOLOv2, is state-of-the-art on standard detection tasks like PASCAL VOC and COCO. Using a novel, multi-scale training method the same YOLOv2 model can r un at varying sizes, offering an easy tradeoff between speed and accuracy. At 67 FPS, YOLOv2 gets 76.8 mAP on VOC 2007. At 40 FPS, YOLOv2 gets 78.6 mAP, outperf orming state-of-the-art methods like Faster RCNN with ResNet and SSD while still running significantly faster. Finally we propose a method to jointly train on o bject detection and classification. Using this method we train YOLO9000 simultan eously on the COCO detection dataset and the ImageNet classification dataset. Ou r joint training allows YOLO9000 to predict detections for object classes that d on't have labelled detection data. We validate our approach on the ImageNet dete ction task. YOLO9000 gets 19.7 mAP on the ImageNet detection validation set desp ite only having detection data for 44 of the 200 classes. On the 156 classes not in COCO, YOLO9000 gets 16.0 mAP. YOLO9000 predicts detections for more than 900 0 different object categories, all in real-time.
**********

Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition
Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, Garrison W. Cottrell; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 , pp. 7272-7281
Recently, there has been a lot of interest in automatically generating descripti ons for an image. Most existing language-model based approaches for this task le arn to generate an image description word by word in its original word order. Ho wever, for humans, it is more natural to locate the objects and their relationsh ips first, and then elaborate on each object, describing notable attributes. We present a coarse-to-fine method that decomposes the original image description i nto a skeleton sentence and its attributes, and generates the skeleton sentence and attribute phrases separately. By this decomposition, our method can generate more accurate and novel descriptions than the previous state-of-the-art. Experi

mental results on the MS-COCO and a larger scale Stock3M datasets show that our algorithm yields consistent improvements across different evaluation metrics, especially on the SPICE metric, which has much higher correlation with human ratings than the conventional metrics. Furthermore, our algorithm can generate descriptions with varied length, benefiting from the separate control of the skeleton and attributes. This enables image description generation that better accommodates user preferences.

**********

A Joint Speaker-Listener-Reinforcer Model for Referring Expressions
Licheng Yu, Hao Tan, Mohit Bansal, Tamara L. Berg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7282-7290
Referring expressions are natural language constructions used to identify particular objects within a scene. In this paper, we propose a unified framework for the tasks of referring expression comprehension and generation. Our model is composed of three modules: speaker, listener, and reinforcer. The speaker generates referring expressions, the listener comprehends referring expressions, and the reinforcer introduces a reward function to guide sampling of more discriminative expressions. The listener-speaker modules are trained jointly in an end-to-end learning framework, allowing the modules to be aware of one another during learning while also benefiting from the discriminative reinforcer's feedback. We demonstrate that this unified framework and training achieves state-of-the-art results for both comprehension and generation on three referring expression datasets. Project and demo page: https://vision.cs.unc.edu/refer

**********

Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields
Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291-7299
We present an approach to efficiently detect the 2D pose of multiple people in an image. The approach uses a nonparametric representation, which we refer to as Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving realtime performance, irrespective of the number of people in the image. The architecture is designed to jointly learn part locations and their association via two branches of the same sequential prediction process. Our method placed first in the inaugural COCO 2016 keypoints challenge, and significantly exceeds the previous state-of-the-art result on the MPII Multi-Person benchmark, both in performance and efficiency.

**********

Newton-Type Methods for Inference in Higher-Order Markov Random Fields
Hariprasad Kannan, Nikos Komodakis, Nikos Paragios; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7300-7309
Linear programming relaxations are central to MAP in- ference in discrete Markov Random Fields. The ability to properly solve the Lagrangian dual is a critical component of such methods. In this paper, we study the benefit of us- ing Newton-type methods to solve the Lagrangian dual of a smooth version of the problem. We investigate their abil- ity to achieve superior convergence behavior and to better handle the ill-conditioned nature of the formulation, as compared to first order methods. We show that it is indeed possible to efficiently apply a trust region Newton method for a broad range of MAP inference problems. In this pa- per we propose a provably globally efficient framework that includes (i) excellent compromise between computational complexity and precision concerning the Hessian matrix construction, (ii) a damping strategy that aids efficient opti- mization, (iii) a truncation strategy coupled with a generic pre-conditioner for Conjugate Gradients, (iv) efficient sum- product computation for sparse clique potentials. Results for higher-order Markov Random Fields demonstrate the potential of this approach.

**********

Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors
Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza

Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, Kevin Murphy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7310-7311

The goal of this paper is to serve as a guide for selecting a detection architecture that achieves the right speed/memory/accuracy balance for a given application and platform. To this end, we investigate various ways to trade accuracy for speed and memory usage in modern convolutional object detection systems. A number of successful systems have been proposed in recent years, but apples-to-apples comparisons are difficult due to different base feature extractors (e.g., VGG, Residual Networks), different default image resolutions, as well as different hardware and software platforms. We present a unified implementation of the Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016) and SSD (Liu et al., 2016) systems, which we view as "meta-architectures" and trace out the speed/accuracy trade-off curve created by using alternative feature extractors and varying other critical parameters such as image size within each of these meta-architectures. On one extreme end of this spectrum where speed and memory are critical, we present a detector that achieves real time speeds and can be deployed on a mobile device. On the opposite end in which accuracy is critical, we present a detector that achieves state-of-the-art performance measured on the COCO detection task.

**********

Deep Outdoor Illumination Estimation

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, Jean-Francois Lalonde; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7312-7321

We present a CNN-based technique to estimate high-dynamic range outdoor illumination from a single low dynamic range image. To train the CNN, we leverage a large dataset of outdoor panoramas. We fit a low-dimensional physically-based outdoor illumination model to the skies in these panoramas giving us a compact set of parameters (including sun position, atmospheric conditions, and camera parameters). We extract limited field-of-view images from the panoramas, and train a CNN with this large set of input image--output lighting parameter pairs. Given a test image, this network can be used to infer illumination parameters that can, in turn, be used to reconstruct an outdoor illumination environment map. We demonstrate that our approach allows the recovery of plausible illumination conditions and enables photorealistic virtual object insertion from a single image. An extensive evaluation on both the panorama dataset and captured HDR environment maps shows that our technique significantly outperforms previous solutions to this problem.

**********

Weakly Supervised Semantic Segmentation Using Web-Crawled Videos

Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, Bohyung Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7322-7330

We propose a novel algorithm for weakly supervised semantic segmentation based on image-level class labels only. In weakly supervised setting, it is commonly observed that trained model overly focuses on discriminative parts rather than the entire object area. Our goal is to overcome this limitation with no additional human intervention by retrieving videos relevant to target class labels from web repository, and generating segmentation labels from the retrieved videos to simulate strong supervision for semantic segmentation. During this process, we take advantage of image classification with discriminative localization technique to reject false alarms in retrieved videos and identify relevant spatio-temporal volumes within retrieved videos. Although the entire procedure does not require any additional supervision, the segmentation annotations obtained from videos are sufficiently strong to learn a model for semantic segmentation. The proposed algorithm substantially outperforms existing methods based on the same level of supervision and is even as competitive as the approaches relying on extra annotations.

**********

Global Optimality in Neural Network Training

Benjamin D. Haeffele, Rene Vidal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7331-7339

The past few years have seen a dramatic increase in the performance of recognition systems thanks to the introduction of deep networks for representation learning. However, the mathematical reasons for this success remain elusive. A key issue is that the neural network training problem is nonconvex, hence optimization algorithms may not return a global minima. This paper provides sufficient conditions to guarantee that local minima are globally optimal and that a local descent strategy can reach a global minima from any initialization. Our conditions require both the network output and the regularization to be positively homogeneous functions of the network parameters, with the regularization being designed to control the network size. Our results apply to networks with one hidden layer, where size is measured by the number of neurons in the hidden layer, and multiple deep subnetworks connected in parallel, where size is measured by the number of subnetworks.

**********

Fine-Tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally

Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, Jianming Liang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7340-7351

Intense interest in applying convolutional neural networks (CNNs) in biomedical image analysis is wide spread, but its success is impeded by the lack of large annotated datasets in biomedical imaging. Annotating biomedical images is not only tedious and time consuming, but also demanding of costly, specialty - oriented knowledge and skills, which are not easily accessible. To dramatically reduce annotation cost, this paper presents a novel method called AIFT (active, incremental fine-tuning) to naturally integrate active learning and transfer learning into a single framework. AIFT starts directly with a pre-trained CNN to seek "worthy" samples from the unannotated for annotation, and the (fine-tuned) CNN is further fine-tuned continuously by incorporating newly annotated samples in each iteration to enhance the CNN's performance incrementally. We have evaluated our method in three different biomedical imaging applications, demonstrating that the cost of annotation can be cut by at least half. This performance is attributed to the several advantages derived from the advanced active and incremental capability of our AIFT method.

**********

FASON: First and Second Order Information Fusion Network for Texture Recognition

Xiyang Dai, Joe Yue-Hei Ng, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7352-7360

Deep networks have shown impressive performance on many computer vision tasks. Recently, deep convolutional neural networks (CNNs) have been used to learn discriminative texture representations. One of the most successful approaches is Bilinear CNN model that explicitly captures the second order statistics within deep features. However, these networks cut off the first order information flow in the deep network and make gradient back-propagation difficult. We propose an effective fusion architecture - FASON that combines second order information flow and first order information flow. Our method allows gradients to back-propagate through both flows freely and can be trained effectively. We then build a multi-level deep architecture to exploit the first and second order information within different convolutional layers. Experiments show that our method achieves improvements over state-of-the-art methods on several benchmark datasets.

**********

Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization

Runpeng Cui, Hu Liu, Changshui Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7361-7369

This work presents a weakly supervised framework with deep neural networks for vision-based continuous sign language recognition, where the ordered gloss labels but no exact temporal locations are available with the video of sign sentence,

and the amount of labeled sentences for training is limited. Our approach addres ses the mapping of video segments to glosses by introducing recurrent convolutio nal neural network for spatio-temporal feature extraction and sequence learning. We design a three-stage optimization process for our architecture. First, we de velop an end-to-end sequence learning scheme and employ connectionist temporal c lassification (CTC) as the objective function for alignment proposal. Second, we take the alignment proposal as stronger supervision to tune our feature extract or. Finally, we optimize the sequence learning model with the improved feature r epresentations, and design a weakly supervised detection network for regularizat ion. We apply the proposed approach to a real-world continuous sign language rec ognition benchmark, and our method, with no extra supervision, achieves results comparable to the state-of-the-art.
**********

On Compressing Deep Models by Low Rank and Sparse Decomposition
Xiyu Yu, Tongliang Liu, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7370-7379
Deep compression refers to removing the redundancy of parameters and feature map s for deep learning models. Low-rank approximation and pruning for sparse struct ures play a vital role in many compression works. However, weight filters tend t o be both low-rank and sparse. Neglecting either part of these structure informa tion in previous methods results in iteratively retraining, compromising accurac y, and low compression rates. Here we propose a unified framework integrating th e low-rank and sparse decomposition of weight matrices with the feature map reco nstructions. Our model includes methods like pruning connections as special case s, and is optimized by a fast SVD-free algorithm. It has been theoretically prov en that, with a small sample, due to its generalizability, our model can well re construct the feature maps on both training and test data, which results in less compromising accuracy prior to the subsequent retraining. With such a "warm sta rt" to retrain, the compression method always possesses several merits: (a) high er compression rates, (b) little loss of accuracy, and (c) fewer rounds to compr ess deep models. The experimental results on several popular models such as Alex Net, VGG-16, and GoogLeNet show that our model can significantly reduce the para meters for both convolutional and fully-connected layers. As a result, our model reduces the size of VGG-16 by 15x, better than other recent compression methods that use a single strategy.
**********

Cross-Modality Binary Code Learning via Fusion Similarity Hashing
Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, Baochang Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp . 7380-7388
Binary code learning has been emerging topic in large-scale cross-modality retri eval recently. It aims to map features from multiple modalities into a common Ha mming space, where the cross-modality similarity can be approximated efficiently via Hamming distance. To this end, most existing works learn binary codes direc tly from data instances in multiple modalities, which preserve both intra- and i nter-modal similarities respectively. Few methods consider to preserve the "fusi on similarity" among multi-modal instances instead, which can explicitly capture their heterogeneous correlation in cross-modality retrieval. In this paper, we propose a hashing scheme, termed Fusion Similarity Hashing (FSH), which explicit ly embeds the graph-based fusion similarity across modalities into a common Hamm ing space. Inspired by the "fusion by diffusion", our core idea is to construct an undirected asymmetric graph to model the fusion similarity among different mo dalities, upon which a graph hashing scheme with alternating optimization is int roduced to learn binary codes that embeds such fusion similarity. Quantitative e valuations on three widely used benchmarks, i.e., UCI Handwritten Digit, MIR-Fli ckr25K and NUS-WIDE, demonstrate that the proposed FSH approach can achieve supe rior performance over the state-of-the-art methods.
**********

Adaptive Relaxed ADMM: Convergence Theory and Practical Implementation
Zheng Xu, Mario A. T. Figueiredo, Xiaoming Yuan, Christoph Studer, Tom Goldstein

; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7389-7398
Many modern computer vision and machine learning applications rely on solving difficult optimization problems that involve non-differentiable objective functions and constraints. The alternating direction method of multipliers (ADMM) is a widely used approach to solve such problems. Relaxed ADMM is a generalization of ADMM that often achieves better performance, but its efficiency depends strongly on algorithm parameters that must be chosen by an expert user. We propose an adaptive method that automatically tunes the key algorithm parameters to achieve optimal performance without user oversight. Inspired by recent work on adaptivity, the proposed adaptive relaxed ADMM (ARADMM) is derived by assuming a Barzilai-Borwein style linear gradient. A detailed convergence analysis of ARADMM is provided, and numerical results on several applications demonstrate fast practical convergence.
**********

Generative Hierarchical Learning of Sparse FRAME Models
Jianwen Xie, Yifei Xu, Erik Nijkamp, Ying Nian Wu, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7399-7407
This paper proposes a method for generative learning of hierarchical random field models. The resulting model, which we call the hierarchical sparse FRAME (Filters, Random field, And Maximum Entropy) model, is a generalization of the original sparse FRAME model by decomposing it into multiple parts that are allowed to shift their locations, scales and rotations, so that the resulting model becomes a hierarchical deformable template. The model can be trained by an EM-type algorithm that alternates the following two steps: (1) Inference: Given the current model, we match it to each training image by inferring the unknown locations, scales, and rotations of the object and its parts by recursive sum-max maps, and (2) Re-learning: Given the inferred geometric configurations of the objects and their parts, we re-learn the model parameters by maximum likelihood estimation via a stochastic gradient algorithm. Experiments show that the proposed method is capable of learning meaningful and interpretable templates that can be used for object detection, classification and clustering.
**********

Exploiting Symmetry and/or Manhattan Properties for 3D Object Structure Estimation From Single and Multiple Images
Yuan Gao, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7408-7417
Many man-made objects have intrinsic symmetries and Manhattan structure. By assuming an orthographic projection model, this paper addresses the estimation of 3D structures and camera projection using symmetry and/or Manhattan structure cues, which occur when the input is single- or multiple-image from the same category, e.g., multiple different cars. Specifically, analysis on the single image case implies that Manhattan alone is sufficient to recover the camera projection, and then the 3D structure can be reconstructed uniquely exploiting symmetry. However, Manhattan structure can be difficult to observe from a single image due to occlusion. To this end, we extend to the multiple-image case which can also exploit symmetry but does not require Manhattan axes. We propose a novel rigid structure from motion method, exploiting symmetry and using multiple images from the same category as input. Experimental results on the Pascal3D+ dataset show that our method significantly outperforms baseline methods.
**********

Fast Haze Removal for Nighttime Image Using Maximum Reflectance Prior
Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, Chang Wen Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7418-7426
In this paper, we address a haze removal problem from a single nighttime image, even in the presence of varicolored and non-uniform illumination. The core idea lies in a novel maximum reflectance prior. We first introduce the nighttime hazy imaging model, which includes a local ambient illumination item in both direct

attenuation term and scattering term. Then, we propose a simple but effective im age prior, maximum reflectance prior, to estimate the varying ambient illuminati on. The maximum reflectance prior is based on a key observation: for most daytim e haze-free image patches, each color channel has very high intensity at some pi xels. For the nighttime haze image, the local maximum intensities at each color channel are mainly contributed by the ambient illumination. Therefore, we can di rectly estimate the ambient illumination and transmission map, and consequently restore a high quality haze-free image. Experimental results on various nighttim e hazy images demonstrate the effectiveness of the proposed approach. In particu lar, our approach has the advantage of computational efficiency, which is 10-100 times faster than state-of-the-art methods.

**********

Adaptive Class Preserving Representation for Image Classification

Jian-Xun Mi, Qiankun Fu, Weisheng Li; Proceedings of the IEEE Conference on Comp uter Vision and Pattern Recognition (CVPR), 2017, pp. 7427-7435

In linear representation-based image classification, an unlabeled sample is repr esented by the entire training set. To obtain a stable and discriminative soluti on, regularization on the vector of representation coefficients is necessary. Fo r example, the representation in sparse representation-based classification (SRC ) uses L1 norm penalty as regularization, which is equal to lasso. However, lass o overemphasizes the role of sparseness while ignoring the inherent structure am ong samples belonging to a same class. Many recent developed representation clas sifications have adopted lasso-type regressions to improve the performance. In t his paper, we propose the adaptive class preserving representation for classific ation (ACPRC). Our method is related to group lasso based classification but dif ferent in two key points: When training samples in a class are uncorrelated, ACP RC turns into SRC; when samples in a class are highly correlated, it obtains sim ilar result as group lasso. The superiority of ACPRC over other state-of-the-art regularization techniques including lasso, group lasso, sparse group lasso, etc . are evaluated by extensive experiments.

**********

A Dataset for Benchmarking Image-Based Localization

Xun Sun, Yuanfan Xie, Pei Luo, Liang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7436-7444

A novel dataset for benchmarking image-based localization is presented. With inc reasing research interests in visual place recognition and localization, several datasets have been published in the past few years. One of the evident limitati ons of existing datasets is that precise ground truth camera poses of query imag es are not available in a meaningful 3D metric system. This is in part due to th e underlying 3D models of these datasets are reconstructed from Structure from M otion methods. So far little attention has been paid to metric evaluations of lo calization accuracy. In this paper we address the problem of whether state-of-th e-art visual localization techniques can be applied to tasks with demanding accu racy requirements. We acquired training data for a large indoor environment with cameras and a LiDAR scanner. In addition, we collected over 2000 query images w ith cell phone cameras. Using LiDAR point clouds as a reference, we employed a s emi-automatic approach to estimate the 6 degrees of freedom camera poses precise ly in the world coordinate system. The proposed dataset enables us to quantitati vely assess the performance of various algorithms using a fair and intuitive met ric.

**********

Low-Rank-Sparse Subspace Representation for Robust Regression

Yongqiang Zhang, Daming Shi, Junbin Gao, Dansong Cheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7445-745 4

Learning robust regression model from high-dimensional corrupted data is an esse ntial and difficult problem in many practical applications. The state-of-the-art methods have studied low-rank regression models that are robust against typical noises (like Gaussian noise and out-sample sparse noise) or outliers, such that a regression model can be learned from clean data lying on underlying subspaces

. However, few of the existing low-rank regression methods can handle the outlie rs/noise lying on the sparsely corrupted disjoint subspaces. To address this iss ue, we propose a low-rank-sparse subspace representation for robust regression, hereafter referred to as LRS-RR in this paper. The main contribution include the following: (1) Unlike most of the existing regression methods, we propose an ap proach with two phases of low-rank-sparse subspace recovery and regression optim ization being carried out simultaneously;(2) we also apply the linearized altern ating direction method with adaptive penalty to solved the formulated LRS-RR pro blem and prove the convergence of the algorithm and analyze its complexity; (3) we demonstrate the efficiency of our method for the high-dimensional corrupted d ata on both synthetic data and two benchmark datasets against several state-of-t he-art robust methods.

**********

AGA: Attribute-Guided Augmentation

Mandar Dixit, Roland Kwitt, Marc Niethammer, Nuno Vasconcelos; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7 455-7463

We consider the problem of data augmentation, i.e., generating artificial sample s to extend a given corpus of training data. Specifically, we propose attributed -guided augmentation (AGA) which learns a mapping that allows to synthesize data such that an attribute of a synthesized sample is at a desired value or strengt h. This is particularly interesting in situations where little data with no attr ibute annotation is available for learning, but we have access to a large extern al corpus of heavily annotated samples. While prior works primarily augment in t he space of images, we propose to perform augmentation in feature space instead. We implement our approach as a deep encoder-decoder architecture that learns t he synthesis function in an end-to-end manner. We demonstrate the utility of our approach on the problems of (1) one-shot object recognition in a transfer-learn ing setting where we have no prior knowledge of the new classes, as well as (2) object-based one-shot scene recognition. As external data, we leverage 3D depth and pose information from the SUN RGB-D dataset. Our experiments show that attri bute-guided augmentation of high-level CNN features considerably improves one-sh ot recognition performance on both problems.

**********

Awesome Typography: Statistics-Based Text Effects Transfer

Shuai Yang, Jiaying Liu, Zhouhui Lian, Zongming Guo; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7464-7473

In this work, we explore the problem of generating fantastic special-effects for the typography. It is quite challenging due to the model diversities to illustr ate varied text effects for different characters. To address this issue, our key idea is to exploit the analytics on the high regularity of the spatial distribu tion for text effects to guide the synthesis process. Specifically, we character ize the stylized patches by their normalized positions and the optimal scales to depict their style elements. Our method first estimates these two features and derives their correlation statistically. They are then converted into soft const raints for texture transfer to accomplish adaptive multi-scale texture synthesis and to make style element distribution uniform. It allows our algorithm to prod uce artistic typography that fits for both local texture patterns and the global spatial distribution in the example. Experimental results demonstrate the super iority of our method for various text effects over conventional style transfer m ethods. In addition, we validate the effectiveness of our algorithm with extensi ve artistic typography library generation.

**********

Lean Crowdsourcing: Combining Humans and Machines in an Online System

Steve Branson, Grant Van Horn, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7474-7483

We introduce a method to greatly reduce the amount of redundant annotations requ ired when crowdsourcing annotations such as bounding boxes, parts, and class lab els. For example, if two Mechanical Turkers happen to click on the same pixel lo cation when annotating a part in a given image--an event that is very unlikely t

o occur by random chance--, it is a strong indication that the location is correct. A similar type of confidence can be obtained if a single Turker happened to agree with a computer vision estimate. We thus incrementally collect a variable number of worker annotations per image based on online estimates of confidence. This is done using a sequential estimation of risk over a probabilistic model that combines worker skill, image difficulty, and an incrementally trained computer vision model. We develop specialized models and algorithms for binary annotation, part keypoint annotation, and sets of bounding box annotations. We show that our method can reduce annotation time by a factor of 4-11 for binary filtering of websearch results, 2-4 for annotation of boxes of pedestrians in images, while in many cases also reducing annotation error. We will make an end-to-end version of our system publicly available.
**********