Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data
Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1-10
While recent deep neural network models have achieved promising results on the image captioning task, they rely largely on the availability of corpora with paired image and sentence captions to describe objects in context. In this work, we propose the Deep Compositional Captioner (DCC) to address the task of generating descriptions of novel objects which are not present in paired image-sentence datasets. Our method achieves this by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. Current deep caption models can only describe objects contained in paired image-sentence corpora, despite the fact that they are pre-trained with large object recognition datasets, namely ImageNet. In contrast, our model can compose sentences that describe novel objects and their interactions with other objects. We demonstrate our model's ability to describe novel concepts by empirically evaluating its performance on MSCOCO and show qualitative results on ImageNet images of objects for which no paired image-caption data exist. Further, we extend our approach to generate descriptions of objects in video clips. Our results show that DCC has distinct advantages over existing image and video captioning approaches for generating descriptions of new objects in context.
****
Generation and Comprehension of Unambiguous Object Descriptions
Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, Kevin Murphy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 11-20
We propose a method that can generate an unambiguous description (known as a referring expression) of a specific object or region in an image, and which can also comprehend or interpret such an expression to infer which object is being described. We show that our method outperforms previous methods that generate descriptions of objects without taking into account other potentially ambiguous objects in the scene. Our model is inspired by recent successes of deep learning methods for image captioning, but while image captioning is difficult to evaluate, our task allows for easy objective evaluation. We also present a new large-scale dataset for referring expressions, based on MS-COCO. We have released the dataset and a toolbox for visualization and evaluation, see https://github.com/mjhucla/Google_Refexp_toolbox.
****
Stacked Attention Networks for Image Question Answering
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29
This paper presents stacked attention networks (SANs)that learn to answer natural language questions from images. SANs use semantic representation of a question as query to search for the regions in an image that are related to the answer. We argue that image question answering (QA) often requires multiple steps of reasoning. Thus, we develop a multiple-layer SAN in which we query an image multiple times to infer the answer progressively. Experiments conducted on four image QA data sets demonstrate that the proposed SANs significantly outperform previous state-of-the-art approaches. The visualization of the attention layers illustrates the progress that the SAN locates the relevant visual clues that lead to the answer of the question layer-by-layer.
****
Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction
Hyeonwoo Noh, Paul Hongsuck Seo, Bohyung Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 30-38
We tackle image question answering (ImageQA) problem by learning a convolutional neural network (CNN) with a dynamic parameter layer whose weights are determine

d adaptively based on questions. For the adaptive parameter prediction, we emplo y a separate parameter prediction network, which consists of gated recurrent uni t (GRU) taking a question as its input and a fully-connected layer generating a set of candidate weights as its output. However, it is challenging to construct a parameter prediction network for a large number of parameters in the fully-con nected dynamic parameter layer of the CNN. We reduce the complexity of this prob lem by incorporating a hashing technique, where the candidate weights given by t he parameter prediction network are selected using a predefined hash function to determine individual weights in the dynamic parameter layer. The proposed netwo rk---joint network with the CNN for ImageQA and the parameter prediction network ---is trained end-to-end through back-propagation, where its weights are initial ized using a pre-trained CNN and GRU. The proposed algorithm illustrates the sta te-of-the-art performance on all available public ImageQA benchmarks.

****

Neural Module Networks

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 39-48

Visual question answering is fundamentally compositional in nature---a question like "where is the dog?" shares substructure with questions like "what color is the dog?" and "where is the cat?" This paper seeks to simultaneously exploit the representational capacity of deep networks and the compositional linguistic str ucture of questions. We describe a procedure for constructing and learning _neu ral module networks_, which compose collections of jointly-trained neural "modul es" into deep networks for question answering. Our approach decomposes questions into their linguistic substructures, and uses these structures to dynamically i nstantiate modular networks (with reusable components for recognizing dogs, clas sifying colors, etc.). The resulting compound networks are jointly trained. We e valuate our approach on two challenging datasets for visual question answering, achieving state-of-the-art results on both the VQA natural image dataset and a n ew dataset of complex questions about abstract shapes.

****

Learning Deep Representations of Fine-Grained Visual Descriptions

Scott Reed, Zeynep Akata, Honglak Lee, Bernt Schiele; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 49-58

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manually-encoded vectors describing shared characteristics among categories. Despite good perform ance, attributes have limitations: (1) finer-grained recognition requires commen surately more attributes, and (2) attributes do not provide a natural language i nterface. We propose to overcome these limitations by training neural language m odels from scratch; i.e. without pre-training and only consuming words and chara cters. Our proposed models train end-to-end to align with the fine-grained and c ategory-specific content of images. Natural language provides a flexible and com pact way of encoding only the salient visual aspects for distinguishing categori es. By training on raw text, our model can do inference on raw text as well, pro viding humans a familiar mode both for annotation and retrieval. Our model achie ves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification o n the Caltech-UCSD Birds 200-2011 dataset.

****

Multi-Cue Zero-Shot Learning With Strong Supervision

Zeynep Akata, Mateusz Malinowski, Mario Fritz, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 59 -68

Scaling up visual category recognition to large numbers of classes remains chall enging. A promising research direction is zero-shot learning, which does not req uire any training data to recognize new classes, but rather relies on some form of auxiliary information describing the new classes. Ultimately, this may allow to use textbook knowledge that humans employ to learn about new classes by trans

ferring knowledge from classes they know well. The most successful zero-shot lea rning approaches currently require a particular type of auxiliary information -- namely attribute annotations performed by humans -- that is not readily availab le for most classes. Our goal is to circumvent this bottleneck by substituting s uch annotations by extracting multiple pieces of information from multiple unstr uctured text sources readily available on the web. To compensate for the weaker form of auxiliary information, we incorporate stronger supervision in the form o f semantic part annotations on the classes from which we transfer knowledge. We achieve our goal by a joint embedding framework that maps multiple text parts as well as multiple semantic parts into a common space. Our results consistently a nd significantly improve on the state-of-the-art in zero-short recognition and r etrieval.
****

Latent Embeddings for Zero-Shot Classification
Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, Bernt Sc hiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni tion (CVPR), 2016, pp. 69-77
We present a novel latent embedding model for learning a compatibility function between image and class embeddings, in the context of zero-shot classification. The proposed method augments the state-of-the-art bilinear compatibility model b y incorporating latent variables. Instead of learning a single bilinear map, it learns a collection of maps with the selection, of which map to use, being a lat ent variable for the current image-class pair. We train the model with a ranking based objective function which penalizes incorrect rankings of the true class f or a given image. We empirically demonstrate that our model improves the state-o f-the-art for various class embeddings consistently on three challenging publicl y available datasets for the zero-shot setting. Moreover, our method leads to vi sually highly interpretable results with clear clusters of different fine-graine d object properties that correspond to different latent variable maps.
****

One-Shot Learning of Scene Locations via Feature Trajectory Transfer
Roland Kwitt, Sebastian Hegenbart, Marc Niethammer; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 78-86
The appearance of (outdoor) scenes changes considerably with the strength of cer tain transient attributes, such as "rainy", "dark" or "sunny". Obviously, this a lso affects the representation of an image in feature space, e.g., as activation s at a certain CNN layer, and consequently impacts scene recognition performance . In this work, we investigate the variability in these transient attributes as a rich source of information for studying how image representations change as a function of attribute strength. In particular, we leverage a recently introduced dataset with fine-grain annotations to estimate feature trajectories for a coll ection of transient attributes and then show how these trajectories can be trans ferred to new image representations. This enables us to synthesize new data alon g the transferred trajectories with respect to the dimensions of the space spann ed by the transient attributes. Applicability of this concept is demonstrated on the problem of one-shot recognition of scene locations. We show that data synth esized via feature trajectory transfer considerably boosts recognition performan ce, (1) with respect to baselines and (2) in combination with state-of-the-art a pproaches in one-shot learning.
****

Learning Attributes Equals Multi-Source Domain Generalization
Chuang Gan, Tianbao Yang, Boqing Gong; Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition (CVPR), 2016, pp. 87-97
Attributes possess appealing properties and benefit many computer vision problem s, such as object recognition, learning with humans in the loop, and image retr ieval. Whereas the existing work mainly pursues utilizing attributes for various computer vision problems, we contend that the most basic problem---how to accur ately and robustly detect attributes from images---has been left under explored. Especially, the existing work rarely explicitly tackles the need that attribute detectors should generalize well across different categories, including those p

reviously unseen. Noting that this is analogous to the objective of multi-source domain generalization, if we treat each category as a domain, we provide a novel perspective to attribute detection and propose to gear the techniques in multi-source domain generalization for the purpose of learning cross-category generalizable attribute detectors. We validate our understanding and approach with extensive experiments on four challenging datasets and three different problems.
****

Anticipating Visual Representations From Unlabeled Video
Carl Vondrick, Hamed Pirsiavash, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 98-106
Anticipating actions and objects before they start or appear is a difficult problem in computer vision with several real-world applications. This task is challenging partly because it requires leveraging extensive knowledge of the world that is difficult to write down. We believe that a promising resource for efficiently learning this knowledge is through readily available unlabeled video. We present a framework that capitalizes on temporal structure in unlabeled video to learn to anticipate human actions and objects. The key idea behind our approach is that we can train deep networks to predict the visual representation of images in the future. Visual representations are a promising prediction target because they encode images at a higher semantic level than pixels yet are automatic to compute. We then apply recognition algorithms on our predicted representation to anticipate objects and actions. We experimentally validate this idea on two datasets, anticipating actions one second in the future and objects five seconds in the future.
****

Learning to Assign Orientations to Feature Points
Kwang Moo Yi, Yannick Verdie, Pascal Fua, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 107-116
We show how to train a Convolutional Neural Network to assign a canonical orientation to feature points given an image patch centered on the feature point. Our method improves feature point matching upon the state-of-the art and can be used in conjunction with any existing rotation sensitive descriptors. To avoid the tedious and almost impossible task of finding a target orientation to learn, we propose to use Siamese networks which implicitly find the optimal orientations during training. We also propose a new type of activation function for Neural Networks that generalizes the popular ReLU, maxout, and PReLU activation functions. This novel activation performs better for our task. We validate the effectiveness of our method extensively with four existing datasets, including two non-planar datasets, as well as our own dataset. We show that we outperform the state-of-the-art without the need of retraining for each dataset.
****

Learning Dense Correspondence via 3D-Guided Cycle Consistency
Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, Alexei A. Efros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 117-126
Discriminative deep learning approaches have shown impressive results for problems where human-labeled ground truth is plentiful, but what about tasks where labels are difficult or impossible to obtain? This paper tackles one such problem: establishing dense visual correspondence across different object instances. For this task, although we do not know what the ground-truth is, we know it should be consistent across instances of that category. We exploit this consistency as a supervisory signal to train a convolutional neural network to predict cross-instance correspondences between pairs of images depicting objects of the same category. For each pair of training images we find an appropriate 3D CAD model and render two synthetic views to link in with the pair, establishing a correspondence flow 4-cycle. We use ground-truth synthetic-to-synthetic correspondences, provided by the rendering engine, to train a ConvNet to predict synthetic-to-real, real-to-real and real-to-synthetic correspondences that are cycle-consistent with the ground-truth. At test time, no CAD models are required. We demonstrate that

our end-to-end trained ConvNet supervised by cycle-consistency outperforms state-of-the-art pairwise matching methods in correspondence-related tasks.
****

The Global Patch Collider
Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, Pushmeet Kohli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 127-135
This paper proposes a novel extremely efficient, fully-parallelizable, task-specific algorithm for the computation of global point-wise correspondences in images and videos. Our algorithm, the Global Patch Collider, is based on detecting unique collisions between image points using a collection of learned tree structures that act as conditional hash functions. In contrast to conventional approaches that rely on pairwise distance computation, our algorithm isolates distinctive pixel pairs that hit the same leaf during traversal through multiple learned tree structures. The split functions stored at the intermediate nodes of the trees are trained to ensure that only visually similar patches or their geometric or photometric transformed versions fall into the same leaf node. The matching process involves passing all pixel positions in the images under analysis through the tree structures. We then compute matches by isolating points that uniquely collide with each other ie. fell in the same empty leaf in multiple trees. Our algorithm is linear in the number of pixels but can be made constant time on a parallel computation architecture as the tree traversal for individual image points is decoupled. We demonstrate the efficacy of our method by using it to perform optical flow matching and stereo matching on some challenging benchmarks. Experimental results show that not only is our method extremely computationally efficient, but it is also able to match or outperform state of the art methods that are much more complex.
****

Joint Probabilistic Matching Using m-Best Solutions
Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 136-145
Matching between two sets of objects is typically approached by finding the object pairs that collectively maximize the joint matching score. In this paper, we argue that this single solution does not necessarily lead to the optimal matching accuracy and that general one-to-one assignment problems can be improved by considering multiple hypotheses before computing the final similarity measure. To that end, we propose to utilize the marginal distributions for each entity. Previously, this idea has been neglected mainly because exact marginalization is intractable due to a combinatorial number of all possible matching permutations. Here, we propose a generic approach to efficiently approximate the marginal distributions by exploiting the m-best solutions of the original problem. This approach not only improves the matching solution, but also provides more accurate ranking of the results, because of the extra information included in the marginal distribution. We validate our claim on two distinct objectives: (i) person re-identification and temporal matching modelled as an integer linear program, and (ii) feature point matching using a quadratic cost function. Our experiments confirm that marginalization indeed leads to superior performance compared to the single (nearly) optimal solution, yielding state-of-the-art results in both applications on standard benchmarks.
****

Face Alignment Across Large Poses: A 3D Solution
Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 146-155
Face alignment, which fits a face model to an image and extracts the semantic meanings of facial pixels, has been an important topic in CV community. However, most algorithms are designed for faces in small to medium poses (below 45 degree), lacking the ability to align faces in large-pose up to 90 degree. The challenges are three-fold: Firstly, the commonly used landmark-based face model assumes

that all the landmarks are visible and is therefore not suitable for profile vie
ws. Secondly, the face appearance varies more dramatically in large poses, rangi
ng from frontal view to profile view. Thirdly, labelling landmarks in large pose
s is an extremely challenging work since the invisible landmarks have to be gues
sed. In this paper, we propose a solution to the three problems in an new alignm
ent framework, called 3D Dense Face Alignment (3DDFA), in which a dense 3D face
model is fitted to the image via convolutional neutral network (CNN). We also pr
opose a method to synthesize large-scale training samples in profile views to so
lve the third problem of data labelling. Experiments on the challenging AFLW dat
abase show that our approach achieves significant improvements over state-of-the
-art methods.
****

Interactive Segmentation on RGBD Images via Cue Selection
Jie Feng, Brian Price, Scott Cohen, Shih-Fu Chang; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 156-164
Interactive image segmentation is an important problem in computer vision with m
any applications including image editing, object recognition and image retrieval
. Most existing interactive segmentation methods only operate on color images. U
ntil recently, very few works have been proposed to leverage depth information f
rom low-cost sensors to improve interactive segmentation. While these methods ac
hieve better results than color-based methods, they are still limited in either
using depth as an additional color channel or simply combining depth with color
in a linear way. We propose a novel interactive segmentation algorithm which can
 incorporate multiple feature cues like color, depth, and normals in an unified
graph cut framework to leverage these cues more effectively. A key contribution
of our method is that it automatically selects a single cue to be used at each p
ixel, based on the intuition that only one cue is necessary to determine the seg
mentation label locally. This is achieved by optimizing over both segmentation
labels and cue labels, using terms designed to decide where both the segmentatio
n and label cues should change. Our algorithm thus produces not only the segment
ation mask but also a cue label map that indicates where each cue contributes to
 the final result. Extensive experiments on five large scale RGBD datasets show
that our proposed algorithm performs significantly better than both other color-
based and RGBD based algorithms in reducing the amount of user inputs as well as
 increasing segmentation accuracy.
****

Layered Scene Decomposition via the Occlusion-CRF
Chen Liu, Pushmeet Kohli, Yasutaka Furukawa; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 165-173
This paper addresses the challenging problem of perceiving the hidden or occlude
d geometry of the scene depicted in any given RGBD image. Unlike other image lab
eling problems such as image segmentation where each pixel needs to be assigned
a single label, layered decomposition requires us to assign multiple labels to p
ixels. We propose a novel "Occlusion-CRF" model that allows for the integration
of sophisticated priors to regularize the solution space and enables the automat
ic inference of the layer decomposition. We use a generalization of the Fusion M
ove algorithm to perform Maximum a Posterior (MAP) inference on the model that c
an handle the large label sets needed to represent multiple surface assignments
to each pixel. We have evaluated the proposed model and the inference algorithm
on many RGBD images of cluttered indoor scenes. Our experiments show that not on
ly is our model able to explain occlusions but it also enables automatic inpaint
ing of occluded/invisible surfaces.
****

Affinity CNN: Learning Pixel-Centric Pairwise Relations for Figure/Ground Embedd
ing
Michael Maire, Takuya Narihira, Stella X. Yu; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 174-182
Spectral embedding provides a framework for solving perceptual organization prob
lems, including image segmentation and figure/ground organization. From an affi
nity matrix describing pairwise relationships between pixels, it clusters pixels

into regions, and, using a complex-valued extension, orders pixels according to layer. We train a convolutional neural network (CNN) to directly predict the pairwise relationships that define this affinity matrix. Spectral embedding then resolves these predictions into a globally-consistent segmentation and figure/ground organization of the scene. Experiments demonstrate significant benefit to this direct coupling compared to prior works which use explicit intermediate stages, such as edge detection, on the pathway from image to affinities. Our results suggest spectral embedding as a powerful alternative to the conditional random field (CRF)-based globalization schemes typically coupled to deep neural networks.

****

Weakly Supervised Object Boundaries

Anna Khoreva, Rodrigo Benenson, Mohamed Omran, Matthias Hein, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 183-192

State-of-the-art learning based boundary detection methods require extensive training data. Since labelling object boundaries is one of the most expensive types of annotations, there is a need to relax the requirement to carefully annotate images to make both the training more affordable and to extend the amount of training data. In this paper we propose a technique to generate weakly supervised annotations and show that bounding box annotations alone suffice to reach high-quality object boundaries without using any object-specific boundary annotations. With the proposed weak supervision techniques we achieve the top performance on the object boundary detection task, outperforming by a large margin the current fully supervised state-of-the-art methods.

****

Object Contour Detection With a Fully Convolutional Encoder-Decoder Network

Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 193-202

We develop a deep learning algorithm for contour detection with a fully convolutional encoder-decoder network. Different from previous low-level edge detection, our algorithm focuses on detecting higher-level object contours. Our network is trained end-to-end on PASCAL VOC with refined ground truth from inaccurate polygon annotations, yielding much higher precision in object contour detection than previous methods. We find that the learned model generalizes well to unseen object classes from the same supercategories on MS COCO and can match state-of-the-art edge detection on BSDS500 with fine-tuning. By combining with the multiscale combinatorial grouping algorithm, our method can generate high-quality segmented object proposals, which significantly advance the state-of-the-art on PASCAL VOC (improving average recall from 0.62 to 0.67) with a relatively small amount of candidates (1660 per image).

****

What Value Do Explicit High Level Concepts Have in Vision to Language Problems?

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 203-212

Much recent progress in Vision-to-Language (V2L) problems has been achieved through a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). This approach does not explicitly represent high-level semantic concepts, but rather seeks to progress directly from image features to text. In this paper we investigate whether this direct approach succeeds due to, or despite, the fact that it avoids the explicit representation of high-level information. We propose a method of incorporating high-level concepts into the successful CNN-RNN approach, and show that it achieves a significant improvement on the state-of-the-art in both image captioning and visual question answering. We also show that the same mechanism can be used to introduce external semantic information and that doing so further improves performance. We achieve the best reported results on both image captioning and VQA on several benchmark datasets, and provide an analysis of the value of explicit high-level concepts in V2L problems.

****
Fast Detection of Curved Edges at Low SNR
Nati Ofir, Meirav Galun, Boaz Nadler, Ronen Basri; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 213-221
Detecting edges is a fundamental problem in computer vision with many applicatio
ns, some involving very noisy images. While most edge detection methods are fast
, they perform well only on relatively clean images. Unfortunately, sophisticate
d methods that are robust to high levels of noise are quite slow. In this paper
we develop a novel multiscale method to detect curved edges in noisy images. Eve
n though our algorithm searches for edges over an exponentially large set of can
didate curves, its runtime is nearly linear in the total number of image pixels.
 As we demonstrate experimentally, our algorithm is orders of magnitude faster t
han previous methods designed to deal with high noise levels. At the same time
it obtains comparable and often superior results to existing methods on a variet
y of challenging noisy images.
****
Object Skeleton Extraction in Natural Images by Fusing Scale-Associated Deep Sid
e Outputs
Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Zhijiang Zhang, Xiang Bai; Proceedings
 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,
 pp. 222-230
Object skeleton is a useful cue for object detection, complementary to the objec
t contour, as it provides a structural representation to describe the relationsh
ip among object parts. While object skeleton extraction in natural images is a v
ery challenging problem, as it requires the extractor to be able to capture both
 local and global image context to determine the intrinsic scale of each skeleto
n pixel. Existing methods rely on per-pixel based multi-scale feature computatio
n, which results in difficult modeling and high time consumption. In this paper,
 we present a fully convolutional network with multiple scale-associated side ou
tputs to address this problem. By observing the relationship between the recepti
ve field sizes of the sequential stages in the network and the skeleton scales t
hey can capture, we introduce a scale-associated side output to each stage. We i
mpose supervision to different stages by guiding the scale-associated side outpu
ts toward groundtruth skeletons of different scales. The responses of the multip
le scale-associated side outputs are then fused in a scale-specific way to local
ize skeleton pixels with multiple scales effectively. Our method achieves promis
ing results on two skeleton extraction datasets, and significantly outperforms o
ther competitors.
****
Learning Relaxed Deep Supervision for Better Edge Detection
Yu Liu, Michael S. Lew; Proceedings of the IEEE Conference on Computer Vision an
d Pattern Recognition (CVPR), 2016, pp. 231-240
We propose using relaxed deep supervision (RDS) within convolutional neural netw
orks for edge detection. The conventional deep supervision utilizes the general
ground-truth to guide intermediate predictions. Instead, we build hierarchical s
upervisory signals with additional relaxed labels to consider the diversities in
 deep neural networks. We begin by capturing the relaxed labels from simple dete
ctors (e.g. Canny).Then we merge them with the general ground-truth to generate
the RDS. Finally we employ the RDS to supervise the edge network following a coa
rse-to-fine paradigm. These relaxed labels can be seen as some false positives t
hat are difficult to be classified. We consider these false positives in the sup
ervision, and are able to achieve high performance for better edge detection. We
 compensate for the lack of training images by capturing coarse edge annotations
 from a large dataset of image segmentations to pretrain the model. Extensive ex
periments demonstrate that our approach achieves state-of-the-art performance on
 the well-known BSDS500 dataset (ODS F-score of .792) and obtains superior cross
-dataset generalization results on NYUD dataset.
****
Occlusion Boundary Detection via Deep Exploration of Context
Huan Fu, Chaohui Wang, Dacheng Tao, Michael J. Black; Proceedings of the IEEE Co

nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 241-250
Occlusion boundaries contain rich perceptual information about the underlying sc
ene structure. They also provide important cues in many visual perception tasks
such as scene understanding, object recognition, and segmentation. In this paper
, we improve occlusion boundary detection via enhanced exploration of contextual
 information (e.g., local structural boundary patterns, observations from surrou
nding regions, and temporal context), and in doing so develop a novel approach b
ased on convolutional neural networks (CNNs) and conditional random fields (CRFs
). Experimental results demonstrate that our detector significantly outperforms
the state-of-the-art (e.g., improving the F-measure from 0.62 to 0.71 on the com
monly used CMU benchmark). Last but not least, we empirically assess the roles o
f several important components of the proposed detector, so as to validate the r
ationale behind this approach.
****

SemiContour: A Semi-Supervised Learning Approach for Contour Detection
Zizhao Zhang, Fuyong Xing, Xiaoshuang Shi, Lin Yang; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 251-259
Supervised contour detection methods usually require many labeled training image
s to obtain satisfactory performance. However, a large set of annotated data mig
ht be unavailable or extremely labor intensive. In this paper, we investigate th
e usage of semi-supervised learning (SSL) to obtain competitive detection accura
cy with very limited training data (three labeled images). Specifically, we prop
ose a semi-supervised structured ensemble learning approach for contour detectio
n built on structured random forests (SRF). To allow SRF to be applicable to unl
abeled data, we present an effective sparse representation approach to capture i
nherent structure in image patches by finding a compact and discriminative low-d
imensional subspace representation in an unsupervised manner, enabling the incor
poration of abundant unlabeled patches with their estimated structured labels to
 help SRF perform better node splitting. We re-examine the role of sparsity and
propose a novel and fast sparse coding algorithm to boost the overall learning e
fficiency. To the best of our knowledge, this is the first attempt to apply SSL
for contour detection. Extensive experiments on the BSDS500 segmentation dataset
 and the NYU Depth dataset demonstrate the superiority of the proposed method.
****

Learning to Localize Little Landmarks
Saurabh Singh, Derek Hoiem, David Forsyth; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2016, pp. 260-269
We interact everyday with tiny objects such as the door handle of a car or the l
ight switch in a room. These little landmarks are barely visible and hard to loc
alize in images. We describe a method to find such landmarks by finding a sequen
ce of latent landmarks, each with a prediction model. Each latent landmark predi
cts the next in sequence, and the last localizes the target landmark. For exampl
e, to find the door handle of a car, our method learns to start with a latent la
ndmark near the wheel, as it is globally distinctive; subsequent latent landmark
s use the context from the earlier ones to get closer to the target. Our method
is supervised solely by the location of the little landmark and displays strong
performance on more difficult variants of established tasks and on two new tasks
.
****

InterActive: Inter-Layer Activeness Propagation
Lingxi Xie, Liang Zheng, Jingdong Wang, Alan L. Yuille, Qi Tian; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.
 270-279
An increasing number of computer vision tasks can be tackled with deep features,
 which are the intermediate outputs of a pre-trained Convolutional Neural Networ
k. Despite the astonishing performance, deep features extracted from low-level n
eurons are still below satisfaction, arguably because they cannot access the spa
tial context contained in the higher layers. In this paper, we present InterActi
ve, a novel algorithm which computes the activeness of neurons and network conne
ctions. Activeness is propagated through a neural network in a top-down manner,

carrying high-level context and improving the descriptive power of low-level and mid-level neurons. Visualization indicates that neuron activeness can be interpreted as spatial-weighted neuron responses. We achieve state-of-the-art classification performance on a wide range of image datasets.
****

Exploit Bounding Box Annotations for Multi-Label Object Recognition
Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, Jianfei Cai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 280-288
Convolutional neural networks (CNNs) have shown great performance as general feature representations for object recognition applications. However, for multi-label images that contain multiple objects from different categories, scales and locations, global CNN features are not optimal. In this paper, we incorporate local information to enhance the feature discriminative power. In particular, we first extract object proposals from each image. With each image treated as a bag and object proposals extracted from it treated as instances, we transform the multi-label recognition problem into a multi-class multi-instance learning problem. Then, in addition to extracting the typical CNN feature representation from each proposal, we propose to make use of ground-truth bounding box annotations (strong labels) to add another level of local information by using nearest-neighbor relationships of local regions to form a multi-view pipeline. The proposed multi-view multi-instance framework utilizes both weak and strong labels effectively, and more importantly it has the generalization ability to even boost the performance of unseen categories by partial strong labels from other categories. Our framework is extensively compared with state-of-the-art hand-crafted feature based methods and CNN based methods on two multi-label benchmark datasets. The experimental results validate the discriminative power and the generalization ability of the proposed framework. With strong labels, our framework is able to achieve state-of-the-art results in both datasets.
****

TI-Pooling: Transformation-Invariant Pooling for Feature Learning in Convolutional Neural Networks
Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 289-297
In this paper we present a deep neural network topology that incorporates a simple to implement transformation-invariant pooling operator (TI-pooling). This operator is able to efficiently handle prior knowledge on nuisance variations in the data, such as rotation or scale changes. Most current methods usually make use of dataset augmentation to address this issue, but this requires larger number of model parameters and more training data, and results in significantly increased training time and larger chance of under- or overfitting. The main reason for these drawbacks is that that the learned model needs to capture adequate features for all the possible transformations of the input. On the other hand, we formulate features in convolutional neural networks to be transformation-invariant. We achieve that using parallel siamese architectures for the considered transformation set and applying the TI-pooling operator on their outputs before the fully-connected layers. We show that this topology internally finds the most optimal "canonical" instance of the input image for training and therefore limits the redundancy in learned features. This more efficient use of training data results in better performance on popular benchmark datasets with smaller number of parameters when comparing to standard convolutional neural networks with dataset augmentation and to other baselines.
****

Fashion Style in 128 Floats: Joint Ranking and Classification Using Weak Data for Feature Extraction
Edgar Simo-Serra, Hiroshi Ishikawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 298-307
We propose a novel approach for learning features from weakly-supervised data by joint ranking and classification. In order to exploit data with weak labels, we

jointly train a feature extraction network with a ranking loss and a classification network with a cross-entropy loss. We obtain high-quality compact discriminative features with few parameters, learned on relatively small datasets without additional annotations. This enables us to tackle tasks with specialized images not very similar to the more generic ones in existing fully-supervised datasets. We show that the resulting features in combination with a linear classifier surpass the state-of-the-art on the Hipster Wars dataset despite using features only 0.3% of the size. Our proposed features significantly outperform those obtained from networks trained on ImageNet, despite being 32 times smaller (128 single-precision floats), trained on noisy and weakly-labeled data, and using only 1.5% of the number of parameters.

****

Equiangular Kernel Dictionary Learning With Applications to Dynamic Texture Analysis

Yuhui Quan, Chenglong Bao, Hui Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 308-316

Most existing dictionary learning algorithms consider a linear sparse model, which often cannot effectively characterize the nonlinear properties present in many types of visual data, e.g. dynamic texture (DT). Such nonlinear properties can be exploited by the so-called kernel sparse coding. This paper proposed an equiangular kernel dictionary learning method with optimal mutual coherence to exploit the nonlinear sparsity of high-dimensional visual data. Two main issues are addressed in the proposed method: (1) coding stability for redundant dictionary of infinite-dimensional space; and (2) computational efficiency for computing kernel matrix of training samples of high-dimensional data. The proposed kernel sparse coding method is applied to dynamic texture analysis with both local DT pattern extraction and global DT pattern characterization. The experimental results showed its performance gain over existing methods.

****

Compact Bilinear Pooling

Yang Gao, Oscar Beijbom, Ning Zhang, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 317-326

Bilinear models has been shown to achieve impressive performance on a wide range of visual tasks, such as semantic segmentation, fine grained recognition and face recognition. However, bilinear features are high dimensional, typically on the order of hundreds of thousands to a few million, which makes them impractical for subsequent analysis. We propose two compact bilinear representations with the same discriminative power as the full bilinear representation but with only a few thousand dimensions. Our compact representations allow back-propagation of classification errors enabling an end-to-end optimization of the visual recognition system. The compact bilinear representations are derived through a novel kernelized analysis of bilinear pooling which provide insights into the discriminative power of bilinear pooling, and a platform for further research in compact pooling methods. Experimentation illustrate the utility of the proposed representations for image classification and few-shot learning across several datasets.

****

Accumulated Stability Voting: A Robust Descriptor From Descriptors of Multiple Scales

Tsun-Yi Yang, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 327-335

This paper proposes a novel local descriptor through accumulated stability voting (ASV). The stability of feature dimensions is measured by their differences across scales. To be more robust to noise, the stability is further quantized by thresholding. The principle of maximum entropy is utilized for determining the best thresholds for maximizing discriminant power of the resultant descriptor. Accumulating stability renders a real-valued descriptor and it can be converted into a binary descriptor by an additional thresholding process. The real-valued descriptor attains high matching accuracy while the binary descriptor makes a good compromise between storage and accuracy. Our descriptors are simple yet effective, and easy to implement. In addition, our descriptors require no training. Expe

riments on popular benchmarks demonstrate the effectiveness of our descriptors a
nd their superiority to the state-of-the-art descriptors.
****

CoMaL: Good Features to Match on Object Boundaries
Swarna K. Ravindran, Anurag Mittal; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2016, pp. 336-345
Traditional Feature Detectors and Trackers use information aggregation in 2D pat
ches to detect and match discriminative patches. However, this information does
not remain the same at object boundaries when there is object motion against a
significantly varying background. In this paper, we propose a new approach for
feature detection, tracking and re-detection that gives significantly improved r
esults at the object boundaries. We utilize level lines or iso-intensity curves
that often remain stable and can be reliably detected even at the object bounda
ries, which they often trace. Stable portions of long level lines are detected a
nd points of high curvature are detected on such curves for corner detection.
Further, this level line is used to separate the portions belonging to the two o
bjects, which is then used for robust matching of such points. While such CoMaL
(Corners on Maximally-stable Level Line Segments) points were found to be much
more reliable at the object boundary regions, they perform comparably at the int
erior regions as well. This is illustrated in exhaustive experiments on real-wor
ld datasets.
****

Progressive Feature Matching With Alternate Descriptor Selection and Corresponde
nce Enrichment
Yuan-Ting Hu, Yen-Yu Lin; Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition (CVPR), 2016, pp. 346-354
We address two difficulties in establishing an accurate system for image matchin
g. First, image matching relies on the descriptor for feature extraction, but th
e optimal descriptor often varies from image to image, or even patch to patch. S
econd, conventional matching approaches carry out geometric checking on a small
set of correspondence candidates due to the concern of efficiency. It may result
in restricted performance in recall. We aim at tackling the two issues by integ
rating adaptive descriptor selection and progressive candidate enrichment into i
mage matching. We consider that the two integrated components are complementary:
The high-quality matching yielded by adaptively selected descriptors helps in e
xploring more plausible candidates, while the enriched candidate set serves as a
better reference for descriptor selection. It motivates us to formulate image m
atching as a joint optimization problem, in which adaptive descriptor selection
and progressive correspondence enrichment are alternately conducted. Our approac
h is comprehensively evaluated and compared with the state-of-the-art approaches
on two benchmarks. The promising results manifest its effectiveness.
****

A New Finsler Minimal Path Model With Curvature Penalization for Image Segmentat
ion and Closed Contour Detection
Da Chen, Jean-Marie Mirebeau, Laurent D. Cohen; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 355-363
In this paper, we propose a new curvature penalized minimal path model for image
segmentation via closed contour detection based on the weighted Euler elastica
curves, firstly introduced to the field of computer vision in [22]. Our image se
gmentation method extracts a collection of curvature penalized minimal geodesics
, concatenated to form a closed contour, by connecting a set of user-specified p
oints. Globally optimal minimal paths can be computed by solving an Eikonal equa
tion. This first order PDE is traditionally regarded as unable to penalize curva
ture, which is related to the path acceleration in active contour models. We int
roduce here a new approach that enables finding a global minimum of the geodesic
energy including a curvature term. We achieve this through the use of a novel F
insler metric adding to the image domain the orientation as an extra space dimen
sion. This metric is non-Riemannian and asymmetric, defined on an orientation li
fted space, incorporating the curvature penalty in the geodesic energy. Experime
nts show that the proposed Finsler minimal path model indeed outperforms state-o

f-the-art minimal path models in both synthetic and real images.
****
Scale-Aware Alignment of Hierarchical Image Segmentation
Yuhua Chen, Dengxin Dai, Jordi Pont-Tuset, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 364-372
Image segmentation is a key component in many computer vision systems, and it is recovering a prominent spot in the literature as methods improve and overcome their limitations. The outputs of most recent algorithms are in the form of a hierarchical segmentation, which provides segmentation at different scales in a single tree-like structure. Commonly, these hierarchical methods start from some low-level features, and are not aware of the scale information of the different regions in them. As such, one might need to work on many different levels of the hierarchy to find the objects in the scene. This work tries to modify the existing hierarchical algorithm by improving their alignment, that is, by trying to modify the depth of the regions in the tree to better couple depth and scale. To do so, we first train a regressor to predict the scale of regions using mid-level features. We then define the anchor slice as the set of regions that better balance between over-segmentation and under-segmentation. The output of our method is an improved hierarchy, re-aligned by the anchor slice. To demonstrate the power of our method, we perform comprehensive experiments, which show that our method, as a post-processing step, can significantly improve the quality of the hierarchical segmentation representations, and ease the usage of hierarchical image segmentation to high-level vision tasks such as object segmentation. We also prove that the improvement generalizes well across different algorithms and datasets, with a low computational cost.
****
Deep Interactive Object Selection
Ning Xu, Brian Price, Scott Cohen, Jimei Yang, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 373-381
Interactive object selection is a very important research problem and has many applications. Previous algorithms require substantial user interactions to estimate the foreground and background distributions. In this paper, we present a novel deep-learning-based algorithm which has much better understanding of objectness and can reduce user interactions to just a few clicks. Our algorithm transforms user-provided positive and negative clicks into two Euclidean distance maps which are then concatenated with the RBG channels of images to compose (image, user interactions) pairs. We generate many of such pairs by combining several random sampling strategies to model users' click patterns and use them to finetune deep Fully Convolutional Networks (FCNs). Finally the output probability maps of our FCN-8s model is integrated with graph cut optimization to refine the boundary segments. Our model is trained on the PASCAL segmentation dataset and evaluated on other datasets with different object classes. Experimental results on both seen and unseen objects clearly demonstrate that our algorithm has a good generalization ability and is superior to all existing interactive object selection approaches.
****
Pull the Plug? Predicting If Computers or Humans Should Segment Images
Danna Gurari, Suyog Jain, Margrit Betke, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 382-391
Foreground object segmentation is a critical step for many image analysis tasks. While automated methods can produce high-quality results, their failures disappoint users in need of practical solutions. We propose a resource allocation framework for predicting how best to allocate a fixed budget of human annotation effort in order to collect higher quality segmentations for a given batch of images and automated methods. The framework is based on a proposed prediction module that estimates the quality of given algorithm-drawn segmentations. We demonstrate the value of the framework for two novel tasks related to "pulling the plug" on computer and human annotators. Specifically, we implement two systems that

automatically decide, for a batch of images, when to replace 1) humans with com
puters to create coarse segmentations required to initialize segmentation tools
and 2) computers with humans to create final, fine-grained segmentations. Exper
iments demonstrate the advantage of relying on a mix of human and computer effor
ts over relying on either resource alone for segmenting objects in three diverse
datasets representing visible, phase contrast microscopy, and fluorescence micr
oscopy images.
****

In the Shadows, Shape Priors Shine: Using Occlusion to Improve Multi-Region Segm
entation
Yuka Kihara, Matvey Soloviev, Tsuhan Chen; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2016, pp. 392-401
We present a new algorithm for multi-region segmentation of 2D images with objec
ts that may partially occlude each other. Our algorithm is based on the observat
ion that human performance on this task is based both on prior knowledge about p
lausible shapes and taking into account the presence of occluding objects whose
shape is already known - once an occluded region is identified, the shape prior
can be used to guess the shape of the missing part. We capture the former aspect
using a deep learning model of shape; for the latter, we simultaneously minimiz
e the energy of all regions and consider only unoccluded pixels for data agreeme
nt. Existing algorithms incorporating object shape priors consider every objec
t separately in turn and can't distinguish genuine deviation from the expected s
hape from parts missing due to occlusion. We show that our method significantly
improves on the performance of a representative algorithm, as evaluated on both
preprocessed natural and synthetic images. Furthermore, on the synthetic images,
we recover the ground truth segmentation with good accuracy.
****

Convexity Shape Constraints for Image Segmentation
Loic A. Royer, David L. Richmond, Carsten Rother, Bjoern Andres, Dagmar Kainmuel
ler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2016, pp. 402-410
Segmenting an image into multiple components is a central task in computer visio
n. In many practical scenarios, prior knowledge about plausible components is av
ailable. Incorporating such prior knowledge into models and algorithms for image
segmentation is highly desirable, yet can be non-trivial. In this work, we intr
oduce a new approach that allows, for the first time, to constrain some or all c
omponents of a segmentation to have convex shapes. Specifically, we extend the M
inimum Cost Multicut Problem by a class of constraints that enforce convexity. T
o solve instances of this NP-hard integer linear program to optimality, we separ
ate the proposed constraints in the branch-and-cut loop of a state-of-the-art IL
P solver. Results on photographs and micrographs demonstrate the effectiveness o
f the approach as well as its advantages over the state-of-the-art heuristic.
****

MCMC Shape Sampling for Image Segmentation With Nonparametric Shape Priors
Ertunc Erdil, Sinan Yildirim, Mujdat Cetin, Tolga Tasdizen; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 411-
419
Segmenting images of low quality or with missing data is a challenging problem.
Integrating statistical prior information about the shapes to be segmented can i
mprove the segmentation results significantly. Most shape-based segmentation alg
orithms optimize an energy functional and find a point estimate for the object t
o be segmented. This does not provide a measure of the degree of confidence in t
hat result, neither does it provide a picture of other probable solutions based
on the data and the priors. With a statistical view, addressing these issues wou
ld involve the problem of characterizing the posterior densities of the shapes o
f the objects to be segmented. For such characterization, we propose a Markov ch
ain Monte Carlo (MCMC) sampling-based image segmentation algorithm that uses sta
tistical shape priors. In addition to better characterization of the statistical
structure of the problem, such an approach would also have the potential to add
ress issues with getting stuck at local optima, suffered by existing shape-based

segmentation methods. Our approach is able to characterize the posterior probability density in the space of shapes through its samples, and to return multiple solutions, potentially from different modes of a multimodal probability density, which would be encountered, e.g., in segmenting objects from multiple shape classes. We present promising results on a variety of data sets. We also provide an extension for segmenting shapes of objects with parts that can go through independent shape variations. This extension involves the use of local shape priors on object parts and provides robustness to limitations in shape training data size.

****

From Noise Modeling to Blind Image Denoising

Fengyuan Zhu, Guangyong Chen, Pheng-Ann Heng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 420-429

Traditional image denoising algorithms always assume the noise to be homogeneous white Gaussian distributed. However, the noise on real images can be much more complex empirically. This paper addresses this problem and proposes a novel blind image denoising algorithm which can cope with real-world noisy images even when the noise model is not provided. It is realized by modeling image noise with mixture of Gaussian distribution (MoG) which can approximate large varieties of continuous distributions. As the number of components for MoG is unknown practically, this work adopts Bayesian nonparametric technique and proposes a novel Low-rank MoG filter (LR-MoG) to recover clean signals (patches) from noisy ones contaminated by MoG noise. Based on LR-MoG, a novel blind image denoising approach is developed. To test the proposed method, this study conducts extensive experiments on synthesis and real images. Our method achieves the state-of-the-art performance consistently.

****

Efficient and Robust Color Consistency for Community Photo Collections

Jaesik Park, Yu-Wing Tai, Sudipta N. Sinha, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 430-438

We present an efficient technique to optimize color consistency of a collection of images depicting a common scene. Our method first recovers sparse pixel correspondences in the input images and stacks them into a matrix with many missing entries. We show that this matrix satisfies a rank two constraint under a simple color correction model. These parameters can be viewed as pseudo white balance and gamma correction parameters for each input image. We present a robust low-rank matrix factorization method to estimate the unknown parameters of this model. Using them, we improve color consistency of the input images or perform color transfer with any input image as the source. Our approach is insensitive to outliers in the pixel correspondences thereby precluding the need for complex pre-processing steps. We demonstrate high-quality color consistency results on large photo collections of popular tourist landmarks and personal photo collections containing images of people.

****

Needle-Match: Reliable Patch Matching Under High Uncertainty

Or Lotan, Michal Irani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 439-448

Reliable patch-matching forms the basis for many algorithms (super-resolution, denoising, inpainting, etc.) However, when the image quality deteriorates (by noise, blur or geometric distortions), the reliability of patch-matching deteriorates as well. Matched patches in the degraded image, do not necessarily imply similarity of the underlying patches in the (unknown) high-quality image. This restricts the applicability of patch-based methods. In this paper we present a patch representation called "Needle", which consists of small multi-scale versions of the patch and its immediate surrounding region. While the patch at the finest image scale is severely degraded, the degradation decreases dramatically in coarser needle scales, revealing reliable information for matching. We show that the Needle is robust to many types of image degradations, leads to matches faithful to the underlying high-quality patches, and to improvement in existing patch-based methods.

****
ReconNet: Non-Iterative Reconstruction of Images From Compressively Sensed Measurements
Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, Amit Ashok; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 449-458
The goal of this paper is to present a non-iterative and more importantly an extremely fast algorithm to reconstruct images from compressively sensed (CS) random measurements. To this end, we propose a novel convolutional neural network (CNN) architecture which takes in CS measurements of an image as input and outputs an intermediate reconstruction. We call this network, ReconNet. The intermediate reconstruction is fed into an off-the-shelf denoiser to obtain the final reconstructed image. On a standard dataset of images we show significant improvements in reconstruction results (both in terms of PSNR and time complexity) over state-of-the-art iterative CS reconstruction algorithms at various measurement rates. Further, through qualitative experiments on real data collected using our block SPC (single pixel camera), we show that our network is highly robust to sensor noise and can recover visually better quality images than competitive algorithms at extremely low sensing rates of 0.1 and 0.04. To demonstrate that our algorithm can recover semantically informative images even at a low measurement rate of 0.01, we present a very robust proof of concept real-time visual tracking application.
****
Soft-Segmentation Guided Object Motion Deblurring
Jinshan Pan, Zhe Hu, Zhixun Su, Hsin-Ying Lee, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 459-468
Object motion blur is a challenging problem as the foreground and the background in the scenes undergo different types of image degradation due to movements in various directions and speed. Most object motion deblurring methods address this problem by segmenting blurred images into regions where different kernels are estimated and applied for restoration. Segmentation on blurred images is difficult due to ambiguous pixels between regions, but it plays an important role for object motion deblurring. To address these problems, we propose a novel model for object motion deblurring. The proposed model is developed based on a maximum a posterior formulation in which soft-segmentation is incorporated for object layer estimation. We propose an efficient algorithm to jointly estimate object segmentation and camera motion where each layer can be deblurred well under the guidance of the soft-segmentation. Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art object motion deblurring methods on challenging scenarios.
****
Two Illuminant Estimation and User Correction Preference
Dongliang Cheng, Abdelrahman Abdelhamed, Brian Price, Scott Cohen, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 469-477
This paper examines the problem of white-balance correction when a scene contains two illuminations. This is a two step process: 1) estimate the two illuminants; and 2) correct the image. Existing methods attempt to estimate a spatially varying illumination map, however, results are error prone and the resulting illumination maps are too low-resolution to be used for proper spatially varying white-balance correction. In addition, the spatially varying nature of these methods make them computationally intensive. We show that this problem can be effectively addressed by not attempting to obtain a spatially varying illumination map, but instead by performing illumination estimation on large sub-regions of the image. Our approach is able to detect when distinct illuminations are present in the image and accurately measure these illuminants. Since our proposed strategy is not suitable for spatially varying image correction, a user study is performed to see if there is a preference for how the image should be corrected when two illuminants are present, but only a global correction can be applied.

The user study shows that when the illuminations are distinct, there is a preference for the outdoor illumination to be corrected resulting in warmer final result. We use these collective findings to demonstrate an effective two illuminant estimation scheme that produces corrected images that users prefer.
****

Deep Contrast Learning for Salient Object Detection
Guanbin Li, Yizhou Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 478-487
Salient object detection has recently witnessed substantial progress due to powerful features extracted using deep convolutional neural networks (CNNs). However, existing CNN-based methods operate at the patch level instead of the pixel level. Resulting saliency maps are typically blurry, especially near the boundary of salient objects. Furthermore, image patches are treated as independent samples even when they are overlapping, giving rise to significant redundancy in computation and storage. In this paper, we propose an end-to-end deep contrast network to overcome the aforementioned limitations. Our deep network consists of two complementary components, a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The first stream directly produces a saliency map with pixel-level accuracy from an input image. The second stream extracts segment-wise features very efficiently, and better models saliency discontinuities along object boundaries. Finally, a fully connected CRF model can be optionally incorporated to improve spatial coherence and contour localization in the fused result from these two streams. Experimental results demonstrate that our deep model significantly improves the state of the art.
****

Multiview Image Completion With Space Structure Propagation
Seung-Hwan Baek, Inchang Choi, Min H. Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 488-496
We present a multiview image completion method that provides geometric consistency among different views by propagating space structures. Since a user specifies the region to be completed in one of multiview photographs casually taken in a scene, the proposed method enables us to complete the set of photographs with geometric consistency by creating or removing structures on the specified region. The proposed method incorporates photographs to estimate dense depth maps. We initially complete color as well as depth from a view, and then facilitate two stages of structure propagation and structure-guided completion. Structure propagation optimizes space topology in the scene across photographs, while structure-guide completion enhances, and completes local image structure of both depth and color in multiple photographs with structural coherence by searching nearest neighbor fields in relevant views. We demonstrate the effectiveness of the proposed method in completing multiview images.
****

Composition-Preserving Deep Photo Aesthetics Assessment
Long Mai, Hailin Jin, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 497-506
Photo aesthetics assessment is challenging. Deep convolutional neural network (ConvNet) methods have recently shown promising results for aesthetics assessment. The performance of these deep ConvNet methods, however, is often compromised by the constraint that the neural network only takes the fixed-size input. To accommodate this requirement, input images need to be transformed via cropping, scaling, or padding, which often damages image composition, reduces image resolution, or causes image distortion, thus compromising the aesthetics of the original images. In this paper, we present a composition-preserving deep ConvNet method that directly learns aesthetics features from the original input images without any image transformations. Specifically, our method adds an adaptive spatial pooling layer upon the regular convolution and pooling layers to directly handle input images with original sizes and aspect ratios. To allow for multi-scale feature extraction, we develop the Multi-Net Adaptive Spatial Pooling ConvNet architecture which consists of multiple sub-networks with different adaptive spatial pooling sizes and leverage a scene-based aggregation layer to effectively combine th

e predictions from multiple sub-networks. Our experiments on the large-scale aes thetics assessment benchmark (AVA) demonstrate that our method can significantly improve the state-of-the-art results in photo aesthetics assessment.
****

Automatic Image Cropping : A Computational Complexity Study
Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, Zhengqin Li; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 507-5 15
Attention based automatic image cropping aims at preserving the most visually im portant region in an image. A common task in this kind of method is to search fo r the smallest rectangle inside which the summed attention is maximized. We demo nstrate that under appropriate formulations, this task can be achieved using eff icient algorithms with low computational complexity. In a practically useful sc enario where the aspect ratio of the cropping rectangle is given, the problem ca n be solved with a computational complexity linear to the number of image pixels . We also study the possibility of multiple rectangle cropping and a new model f acilitating fully automated image cropping.
****

A Deeper Look at Saliency: Feature Contrast, Semantics, and Beyond
Neil D. B. Bruce, Christopher Catton, Sasa Janjic; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 516-524
In this paper we consider the problem of visual saliency modeling, including bot h human gaze prediction and salient object segmentation. The overarching goal of the paper is to identify high level considerations relevant to deriving more so phisticated visual saliency models. A deep learning model based on fully convolu tional networks (FCNs) is presented, which shows very favorable performance acro ss a wide variety of benchmarks relative to existing proposals. We also demonstr ate that the manner in which training data is selected, and ground truth treated is critical to resulting model behaviour. Recent efforts have explored the rela tionship between human gaze and salient objects, and we also examine this point further in the context of FCNs. Close examination of the proposed and alternativ e models serves as a vehicle for identifying problems important to developing mo re comprehensive models going forward.
****

Spatially Binned ROC: A Comprehensive Saliency Metric
Calden Wloka, John Tsotsos; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2016, pp. 525-534
A recent trend in saliency algorithm development is large-scale benchmarking and algorithm ranking with ground truth provided by datasets of human fixations. In order to accommodate the strong bias humans have toward central fixations, it i s common to replace traditional ROC metrics with a shuffled ROC metric which use s randomly sampled fixations from other images in the database as the negative s et. However, the shuffled ROC introduces a number of problematic elements, inclu ding a fundamental assumption that it is possible to separate visual salience an d image spatial arrangement. We argue that it is more informative to directly m easure the effect of spatial bias on algorithm performance rather than try to co rrect for it. To capture and quantify these known sources of bias, we propose a novel metric for measuring saliency algorithm performance: the spatially binned ROC (spROC). This metric provides direct insight into the spatial biases of a sa liency algorithm without sacrificing the intuitive raw performance evaluation of traditional ROC measurements. By quantitatively measuring the bias in saliency algorithms, researchers will be better equipped to select and optimize the most appropriate algorithm for a given task. We use a baseline measure of inherent al gorithm bias to show that Adaptive Whitening Saliency (AWS) [14], Attention by I nformation Maximization (AIM) [8], and Dynamic Visual Attention (DVA) [20] provi de the least spatially biased results, suiting them for tasks in which there is no information about the underlying spatial bias of the stimuli, whereas algorit hms such as Graph Based Visual Saliency (GBVS) [18] and Context-Aware Saliency (CAS) [15] have a significant inherent central bias.
****

GraB: Visual Saliency via Novel Graph Model and Background Priors
Qiaosong Wang, Wen Zheng, Robinson Piramuthu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 535-543
We propose an unsupervised bottom-up saliency detection approach by exploiting novel graph structure and background priors. The input image is represented as an undirected graph with superpixels as nodes. Feature vectors are extracted from each node to cover regional color, contrast and texture information. A novel graph model is proposed to effectively capture local and global saliency cues. To obtain more accurate saliency estimations, we optimize the saliency map by using a robust background measure. Comprehensive evaluations on benchmark datasets indicate that our algorithm universally surpasses state-of-the-art unsupervised solutions and performs favorably against supervised approaches.
****

Predicting When Saliency Maps Are Accurate and Eye Fixations Consistent
Anna Volokitin, Michael Gygli, Xavier Boix; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 544-552
Many computational models of visual attention use image features and machine learning techniques to predict eye fixation locations as saliency maps. Recently, the success of Deep Convolutional Neural Networks (DCNNs) for object recognition has opened a new avenue for computational models of visual attention due to the tight link between visual attention and object recognition. In this paper, we show that using features from DCNNs for object recognition we can make predictions that enrich the information provided by saliency models. Namely, we can estimate the reliability of a saliency model from the raw image, which serves as a meta-saliency measure that may be used to select the best saliency algorithm for an image. Analogously, the consistency of the eye fixations among subjects, i.e. the agreement between the eye fixation locations of different subjects, can also be predicted and used by a designer to assess whether subjects reach a consensus about salient image locations.
****

Split and Match: Example-Based Adaptive Patch Sampling for Unsupervised Style Transfer
Oriel Frigo, Neus Sabater, Julie Delon, Pierre Hellier; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 553-561
This paper presents a novel unsupervised method to transfer the style of an example image to a source image. The complex notion of image style is here considered as a local texture transfer, eventually coupled with a global color transfer. For the local texture transfer, we propose a new method based on an adaptive patch partition that captures the style of the example image and preserves the structure of the source image. More precisely, this example-based partition predicts how well a source patch matches an example patch. Results on various images show that our method outperforms the most recent techniques.
****

Detection and Accurate Localization of Circular Fiducials Under Highly Challenging Conditions
Lilian Calvet, Pierre Gurdjos, Carsten Griwodz, Simone Gasparini; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 562-570
Using fiducial markers ensures reliable detection and identification of planar features in images. Fiducials are used in a wide range of applications, especially when a reliable visual reference is needed, e.g., to track the camera in cluttered or textureless environments. A marker designed for such applications must be robust to partial occlusions, varying distances and angles of view, and fast camera motions. In this paper, we present a robust, highly accurate fiducial system, whose markers consist of concentric rings, along with its theoretical foundations. Relying on projective properties, it allows to robustly localize the imaged marker and to accurately detect the position of the image of the (common) circle center. We demonstrate that our system can detect and accurately localize these circular fiducials under very challenging conditions and the experimental results reveal that it outperforms other recent fiducial systems.

****
Scene Recognition With CNNs: Objects, Scales and Dataset Bias
Luis Herranz, Shuqiang Jiang, Xiangyang Li; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 571-579
Since scenes are composed in part of objects, accurate recognition of scenes req
uires knowledge about both scenes and objects. In this paper we address two rela
ted problems: 1) scale induced dataset bias in multi-scale convolutional neural
network (CNN) architectures, and 2) how to combine effectively scene-centric and
 object-centric knowledge (i.e. Places and ImageNet) in CNNs. An earlier attempt
, Hybrid-CNN, showed that incorporating ImageNet did not help much. Here we prop
ose an alternative method taking the scale into account, resulting in significan
t recognition gains. By analyzing the response of ImageNet-CNNs and Places-CNNs
at different scales we find that both operate in different scale ranges, so usin
g the same network for all the scales induces dataset bias resulting in limited
performance. Thus, adapting the feature extractor to each particular scale (i.e.
 scale-specific CNNs) is crucial to improve recognition, since the objects in th
e scenes have their specific range of scales. Experimental results show that the
 recognition accuracy highly depends on the scale, and that simple yet carefully
 chosen multi-scale combinations of ImageNet-CNNs and Places-CNNs, can push the
state-of-the-art recognition accuracy in SUN397 up to 66.26% (and even 70.17% wi
th deeper architectures, comparable to human performance).
****
Learning Action Maps of Large Environments via First-Person Vision
Nicholas Rhinehart, Kris M. Kitani; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2016, pp. 580-588
When people observe and interact with physical spaces, they are able to associat
e functionality to regions in the environment. Our goal is to automate functiona
l understanding of large spaces by leveraging activity demonstrations recorded f
rom an ego-centric viewpoint. The method we describe enables functionality estim
ation in both large scenes where people have behaved, as well as novel scenes wh
ere no behaviors are available. Our method learns and predicts "Action Maps", wh
ich encode the ability for a user to perform activities at various locations. Wi
th the usage of an egocentric camera to observe demonstrations, our method scale
s with the size of the scene without the need for mounting multiple static surve
illance cameras, and is well-suited to the task of observing activities up-close
. We demonstrate that by capturing appearance-based attributes of the environmen
t and associating these attributes with activity demonstrations, our mathematica
l framework allows for the prediction of Action Maps in new environments. Additi
onally, we take a preliminary look at the breadth of applicability of Action Map
s by demonstrating a proof-of-concept application in which they are used in conc
ert with activity detections to perform localization.
****
Single-Image Crowd Counting via Multi-Column Convolutional Neural Network
Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, Yi Ma; Proceedings of the
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589
-597
This paper aims to develop a method that can accurately estimate the crowd count
 from an individual image with arbitrary crowd density and arbitrary perspective
. To this end,we have proposed a simple but effective Multi-column Convolutional
 Neural Network (MCNN) architecture to map the image to its crowd density map. T
he proposed MCNN allows the input image to be of arbitrary size or resolution. B
y utilizing filters with receptive fields of different sizes, the features learn
ed by each column CNN are adaptive to variations in people/head size due to pers
pective effect or image resolution. Furthermore, the true density map is compute
d accurately based on geometry-adaptive kernels which do not need knowing the pe
rspective map of the input image. Since exiting crowd counting datasets do not a
dequately cover all the challenging situations considered in our work, we have c
ollected and labelled a large new dataset that includes 1198 images with about 3
30,000 heads annotated. On this challenging new dataset, as well as all existing
 datasets, we conduct extensive experiments to verify the effectiveness of the p

roposed model and method. In particular, with the proposed simple MCNN model, ou r method outperforms all existing methods. In addition, experiments show that ou r model, once trained on one dataset, can be readily transferred to a new datase t.
****

Shallow and Deep Convolutional Networks for Saliency Prediction
Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness, Noel E. O'Conn or; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio n (CVPR), 2016, pp. 598-606
The prediction of salient areas in images has been traditionally addressed with hand-crafted features based on neuroscience principles. This paper, however, add resses the problem with a completely data-driven approach by training a convolut ional neural network (convnet). The learning process is formulated as a minimiza tion of a loss function that measures the Euclidean distance of the predicted sa liency map with the provided ground truth. The recent publication of large datas ets of saliency prediction has provided enough data to train end-to-end architec tures that are both fast and accurate. Two designs are proposed: a shallow convn et trained from scratch, and a another deeper solution whose first three layers are adapted from another network trained for classification. To the authors know ledge, these are the first end-to-end CNNs trained and tested for the purpose of saliency prediction.
****

Sample and Filter: Nonparametric Scene Parsing via Efficient Filtering
Mohammad Najafi, Sarah Taghavi Namin, Mathieu Salzmann, Lars Petersson; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 16, pp. 607-615
Scene parsing has attracted a lot of attention in computer vision. While paramet ric models have proven effective for this task, they cannot easily incorporate n ew training data. By contrast, nonparametric approaches, which bypass any learni ng phase and directly transfer the labels from the training data to the query im ages, can readily exploit new labeled samples as they become available. Unfortun ately, because of the computational cost of their label transfer procedures, sta te-of-the-art nonparametric methods typically filter out most training images to only keep a few relevant ones to label the query. As such, these methods throw away many images that still contain valuable information and generally obtain a n unbalanced set of labeled samples. In this paper, we introduce a nonparametric approach to scene parsing that follows a sample-and-filter strategy. More speci fically, we propose to sample labeled superpixels according to an image similari ty score, which allows us to obtain a balanced set of samples. We then formulate label transfer as an efficient filtering procedure, which lets us exploit more labeled samples than existing techniques. Our experiments evidence the benefits of our approach over state-of-the-art nonparametric methods on two benchmark dat asets.
****

DeLay: Robust Spatial Layout Estimation for Cluttered Indoor Scenes
Saumitro Dasgupta, Kuan Fang, Kevin Chen, Silvio Savarese; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 616-6 24
We consider the problem of estimating the spatial layout of an indoor scene from a monocular RGB image, modeled as the projection of a 3D cuboid. Existing solut ions to this problem often rely strongly on hand-engineered features and vanishi ng point detection, which are prone to failure in the presence of clutter. In th is paper, we present a method that uses a fully convolutional neural network (FC NN) in conjunction with a novel optimization framework for generating layout est imates. We demonstrate that our method is robust in the presence of clutter and handles a wide range of highly challenging scenes. We evaluate our method on two standard benchmarks and show that it achieves state of the art results, outperf orming previous methods by a wide margin.
****

A Text Detection System for Natural Scenes With Convolutional Feature Learning a

nd Cascaded Classification
Siyu Zhu, Richard Zanibbi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 625-632
We propose a system that finds text in natural scenes using a variety of cues. Our novel data-driven method incorporates coarse-to-fine detection of character pixels using convolutional features (Text-Conv), followed by extracting connected components (CCs) from characters using edge and color features, and finally performing a graph-based segmentation of CCs into words (Word-Graph). For Text-Conv, the initial detection is based on convolutional feature maps similar to those used in Convolutional Neural Networks (CNNs), but learned using Convolutional k-means. Convolution masks defined by local and neighboring patch features are used to improve detection accuracy. The Word-Graph algorithm uses contextual information to both improve word segmentation and prune false character/word detections. Different definitions for foreground (text) regions are used to train the detection stages, some based on bounding box intersection, and others on bounding box and pixel intersection. Our system obtains pixel, character, and word detection f-measures of 93.14%, 90.26%, and 86.77% respectively for the ICDAR 2015 Robust Reading Focused Scene Text dataset, out-performing state-of-the-art systems. This approach may work for other detection targets with homogenous color in natural scenes.
****
Reversible Recursive Instance-Level Object Segmentation
Xiaodan Liang, Yunchao Wei, Xiaohui Shen, Zequn Jie, Jiashi Feng, Liang Lin, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 633-641
In this work, we propose a novel Reversible Recursive Instance-level Object Segmentation (R2-IOS) framework to address the challenging instance-level object segmentation task. R2-IOS consists of a reversible proposal refinement sub-network that predicts bounding box offsets for refining the object proposal locations, and an instance-level segmentation sub-network that generates the foreground mask of the dominant object instance in each proposal. By being recursive, R2-IOS iteratively optimizes the two sub-networks during joint training, in which the refined object proposals and improved segmentation predictions are alternately fed into each other to progressively increase the network capabilities. By being reversible, the proposal refinement sub-network adaptively determines an optimal number of refinement iterations required for each proposal during both training and testing. Furthermore, to handle multiple overlapped instances within a proposal, an instance-aware denoising autoencoder is introduced into the segmentation sub-network to distinguish the dominant object from other distracting instances. Extensive experiments on the challenging PASCAL VOC 2012 benchmark well demonstrate the superiority of R2-IOS over other state-of-the-art methods. In particular, the AP^r over 20 classes at 0.5 IoU achieves 66.7%, which significantly outperforms the results of 58.7% by PFN[15] and 46.3% by[17].
****
Coherent Parametric Contours for Interactive Video Object Segmentation
Yao Lu, Xue Bai, Linda Shapiro, Jue Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 642-650
Interactive video segmentation systems aim at producing sub-pixel-level object boundaries for visual effect applications. Recent approaches mainly focus on using sparse user input (i.e. scribbles) for efficient segmentation; however, the quality of the final object boundaries is not satisfactory for the following reasons: (1) the boundary on each frame is often not accurate; (2) boundaries across adjacent frames wiggle around inconsistently, causing temporal flickering; and (3) there is a lack of direct user control for fine tuning. We propose Coherent Parametric Contours, a novel video segmentation propagation framework that addresses all the above issues. Our approach directly models the object boundary using a set of parametric curves, providing direct user controls for manual adjustment. A spatio-temporal optimization algorithm is employed to produce object boundaries that are spatially accurate and temporally stable. We show that existing evaluation datasets are limited and demonstrate a new set to cover the common ca

ses in professional rotoscoping. A new metric for evaluating temporal consistenc y is proposed. Results show that our approach generates higher quality, more coh erent segmentation results than previous methods.
****

Manifold SLIC: A Fast Method to Compute Content-Sensitive Superpixels
Yong-Jin Liu, Cheng-Chi Yu, Min-Jing Yu, Ying He; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 651-659
Superpixels are perceptually meaningful atomic regions that can effectively capt ure image features. Among various methods for computing uniform superpixels, sim ple linear iterative clustering (SLIC) is popular due to its simplicity and high performance. In this paper, we extend SLIC to compute content-sensitive superpi xels, i.e., small superpixels in content-dense regions (e.g., with high intensit y or color variation) and large superpixels in content-sparse regions. Rather th an the conventional SLIC method that clusters pixels in R5, we map the image I t o a 2-dimensional manifold M in R5, whose area elements are a good measure of th e content density in I. We propose an efficient method to compute restricted cen troidal Voronoi tessellation (RCVT) --- a uniform tessellation --- on M, which i nduces the content-sensitive superpixels in I. Unlike other algorithms that char acterize content-sensitivity by geodesic distances, manifold SLIC tackles the pr oblem by measuring areas of Voronoi cells on M, which can be computed at a very low cost. As a result, it runs 10 times faster than the state-of-the-art content -sensitive superpixels algorithm. We evaluate manifold SLIC and seven representa tive methods on the BSDS500 benchmark and observe that our method outperforms th e existing methods.
****

Deep Saliency With Encoded Low Level Distance Map and High Level Features
Gayoung Lee, Yu-Wing Tai, Junmo Kim; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2016, pp. 660-668
Recent advances in saliency detection have utilized deep learning to obtain high level features to detect salient regions in a scene. They have demonstrated sup erior results over previous works that utilize hand-crafted low level features f or saliency detection. In this paper, we demonstrate that the hand-crafted featu res can provide complementary effects to enhance performance of saliency detecti on that utilizes only high level features. Our method utilizes both high level a nd low level features for saliency detection under a unified deep learning frame work. The high level features are extracted using the VGG-net, and the low level features are compared with other parts of an image to form a low level distance map. The low level distance map is then encoded using a CNN with multiple 1*1 c onvolutional and ReLU layers. We concatenate the encoded low level distance map and the high level features, and connect them to a fully connected neural networ k classifier to evaluate the saliency of a query region. Our experiments show th at our method can further improve performance of the state-of-the-art deep learn ing based saliency detection methods.
****

Instance-Level Segmentation for Autonomous Driving With Deep Densely Connected M RFs
Ziyu Zhang, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 669-677
Our aim is to provide a pixel-wise instance-level labeling of a monocular image in the context of autonomous driving. We build on recent work [Zhang et al., ICC V15] that trained a convolutional neural net to predict instance labeling in loc al image patches, extracted exhaustively in a stride from an image. A simple Ma rkov random field model using several heuristics was then proposed in [Zhang et al., ICCV15] to derive a globally consistent instance labeling of the image. In this paper, we formulate the global labeling problem with a novel densely connec ted Markov random field and show how to encode various intuitive potentials in a way that is amenable to efficient mean field inference [Krahenbuhl et al., NI PS11]. Our potentials encode the compatibility between the global labeling and t he patch-level predictions, contrast-sensitive smoothness as well as the fact th at separate regions form different instances. Our experiments on the challenging

KITTI benchmark [Geiger et al., CVPR12] demonstrate that our method achieves a significant performance boost over the baseline [Zhang et al., ICCV15].
****

DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection
Nian Liu, Junwei Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 678-686
Traditional1 salient object detection models often use hand-crafted features to formulate contrast and various prior knowledge, and then combine them artificially. In this work, we propose a novel end-to-end deep hierarchical saliency network (DHSNet) based on convolutional neural networks for detecting salient objects. DHSNet first makes a coarse global prediction by automatically learning various global structured saliency cues, including global contrast, objectness, compactness, and their optimal combination. Then a novel hierarchical recurrent convolutional neural network (HRCNN) is adopted to further hierarchically and progressively refine the details of saliency maps step by step via integrating local context information. The whole architecture works in a global to local and coarse to fine manner. DHSNet is directly trained using whole images and corresponding ground truth saliency masks. When testing, saliency maps can be generated by directly and efficiently feedforwarding testing images through the network, without relying on any other techniques. Evaluations on four benchmark datasets and comparisons with other 11 state-of-the-art algorithms demonstrate that DHSNet not only shows its significant superiority in terms of performance, but also achieves a real-time speed of 23 FPS on modern GPUs.
****

Object Co-Segmentation via Graph Optimized-Flexible Manifold Ranking
Rong Quan, Junwei Han, Dingwen Zhang, Feiping Nie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 687-695
Aiming at automatically discovering the common objects contained in a set of relevant images and segmenting them as foreground simultaneously, object co-segmentation has become an active research topic in recent years. Although a number of approaches have been proposed to address this problem, many of them are designed with the misleading assumption, unscalable prior, or low flexibility and thus still suffer from certain limitations, which reduces their capability in the real-world scenarios. To alleviate these limitations, we propose a novel two-stage co-segmentation framework, which introduces the weak background prior to establish a globally close- loop graph to represent the common object and union background separately. Then a novel graph optimized-flexible manifold ranking algorithm is proposed to flexibly optimize the graph connection and node labels to co-segment the common objects. Experiments on three image datasets demonstrate that our method outperforms other state-of-the-art methods.
****

Primary Object Segmentation in Videos via Alternate Convex Optimization of Foreground and Background Distributions
Won-Dong Jang, Chulwoo Lee, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 696-704
An unsupervised video object segmentation algorithm, which discovers a primary object in a video sequence automatically, is proposed in this work. We introduce three energies in terms of foreground and background probability distributions: Markov, spatiotemporal, and antagonistic energies. Then, we minimize a hybrid of the three energies to separate a primary object from its background. However, the hybrid energy is nonconvex. Therefore, we develop the alternate convex optimization (ACO) scheme, which decomposes the nonconvex optimization into two quadratic programs. Moreover, we propose the forward-backward strategy, which performs the segmentation sequentially from the first to the last frames and then vice versa, to exploit temporal correlations. Experimental results on extensive datasets demonstrate that the proposed ACO algorithm outperforms the state-of-the-art techniques significantly.
****

Automatic Fence Segmentation in Videos of Dynamic Scenes
Renjiao Yi, Jue Wang, Ping Tan; Proceedings of the IEEE Conference on Computer V

ision and Pattern Recognition (CVPR), 2016, pp. 705-713
We present a fully automatic approach to detect and segment fence-like occluders from a video clip. Unlike previous approaches that usually assume either static scenes or cameras, our method is capable of handling both dynamic scenes and moving cameras. Under a bottom-up framework, it first clusters pixels into coherent groups using color and motion features. These pixel groups are then analyzed in a fully connected graph, and labeled as either fence or non-fence using graph-cut optimization. Finally, we solve a dense Conditional Random Filed (CRF) constructed from multiple frames to enhance both spatial accuracy and temporal coherence of the segmentation. Once segmented, one can use existing hole-filling methods to generate a fence-free output. Extensive evaluation suggests that our method outperforms previous automatic and interactive approaches on complex examples captured by mobile devices.
****

Discovering the Physical Parts of an Articulated Object Class From Multiple Videos
Luca Del Pero, Susanna Ricco, Rahul Sukthankar, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 714-723
We propose a motion-based method to discover the physical parts of an articulated object class (e.g. head/torso/leg of a horse) from multiple videos. The key is to find object regions that exhibit consistent motion relative to the rest of the object, across multiple videos. We can then learn a location model for the parts and segment them accurately in the individual videos using an energy function that also enforces temporal and spatial consistency in part motion. Unlike our approach, traditional methods for motion segmentation or non-rigid structure from motion operate on one video at a time. Hence they cannot discover a part unless it displays independent motion in that particular video. We evaluate our method on a new dataset of 32 videos of tigers and horses, where we significantly outperform a recent motion segmentation method on the task of part discovery (obtaining roughly twice the accuracy).
****

A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation
Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, Alexander Sorkine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 724-732
Over the years, datasets and benchmarks have proven their fundamental importance in computer vision research, enabling targeted progress and objective comparisons in many fields. At the same time, legacy datasets may impend the evolution of a field due to saturated algorithm performance and the lack of contemporary, high quality data. In this work we present a new benchmark dataset and evaluation methodology for the area of video object segmentation. The dataset, named DAVIS (Densely Annotated VIdeo Segmentation), consists of fifty high quality, Full HD video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motion-blur and appearance changes. Each video is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. In addition, we provide a comprehensive analysis of several state-of-the-art segmentation approaches using three complementary metrics that measure the spatial extent of the segmentation, the accuracy of the silhouette contours and the temporal coherence. The results uncover strengths and weaknesses of current approaches, opening up promising directions for future works.
****

Learning Temporal Regularity in Video Sequences
Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 733-742
Perceiving meaningful activities in a long video sequence is a challenging problem due to ambiguous definition of `meaningfulness' as well as clutters in the scene. We approach this problem by learning a generative model for regular motion patterns (termed as regularity) using multiple sources with very limited supervi

sion. Specifically, we propose two methods that are built upon the autoencoders for their ability to work with little to no supervision. We first leverage the conventional handcrafted spatio-temporal local features and learn a fully connected autoencoder on them. Second, we build a fully convolutional feed-forward autoencoder to learn both the local features and the classifiers as an end-to-end learning framework. Our model can capture the regularities from multiple datasets. We evaluate our methods in both qualitative and quantitative ways - showing the learned regularity of videos in various aspects and demonstrating competitive performance on anomaly detection datasets as an application.
****

Bilateral Space Video Segmentation
Nicolas Maerki, Federico Perazzi, Oliver Wang, Alexander Sorkine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 743-751
In this work, we propose a novel approach to video segmentation that operates in bilateral space. We design a new energy on the vertices of a regularly sampled spatio-temporal bilateral grid, which can be solved efficiently using a standard graph cut label assignment. Using a bilateral formulation, the energy that we minimize implicitly approximates long-range, spatio-temporal connections between pixels while still containing only a small number of variables and only local graph edges. We compare to a number of recent methods, and show that our approach achieves state-of-the-art results on multiple benchmarks in a fraction of the runtime. Furthermore, our method scales linearly with image size, allowing for interactive feedback on real-world high resolution video.
****

ReD-SFA: Relation Discovery Based Slow Feature Analysis for Trajectory Clustering
Zhang Zhang, Kaiqi Huang, Tieniu Tan, Peipei Yang, Jun Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 752-760
For spectral embedding/clustering, it is still an open problem on how to construct an relation graph to reflect the intrinsic structures in data. In this paper, we proposed an approach, named Relation Discovery based Slow Feature Analysis (ReD-SFA), for feature learning and graph construction simultaneously. Given an initial graph with only a few nearest but most reliable pairwise relations, new reliable relations are discovered by an assumption of reliability preservation, i.e., the reliable relations will preserve their reliabilities in the learnt projection subspace. We formulate the idea as a cross entropy (CE) minimization problem to reduce the discrepancy between two Bernoulli distributions parameterized by the updated distances and the existing relation graph respectively. Furthermore, to overcome the imbalanced distribution of samples, a Boosting-like strategy is proposed to balance the discovered relations over all clusters. To evaluate the proposed method, extensive experiments are performed with various trajectory clustering tasks, including motion segmentation, time series clustering and crowd detection. The results demonstrate that ReD-SFA can discover reliable intra-cluster relations with high precision, and competitive clustering performance can be achieved in comparison with state-of-the-art.
****

Training Region-Based Object Detectors With Online Hard Example Mining
Abhinav Shrivastava, Abhinav Gupta, Ross Girshick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 761-769
The field of object detection has made significant advances riding on the wave of region-based ConvNets, but their training procedure still includes many heuristics and hyperparameters that are costly to tune. We present a simple yet surprisingly effective online hard example mining (OHEM) algorithm for training region-based ConvNet detectors. Our motivation is the same as it has always been -- detection datasets contain an overwhelming number of easy examples and a small number of hard examples. Automatic selection of these hard examples can make training more effective and efficient. OHEM is a simple and intuitive algorithm that eliminates several heuristics and hyperparameters in common use. But more importa

ntly, it yields consistent and significant boosts in detection performance on benchmarks like PASCAL VOC 2007 and 2012. Its effectiveness increases as datasets become larger and more difficult, as demonstrated by the results on the MS COCO dataset. Moreover, combined with complementary advances in the field, OHEM leads to state-of-the-art results of 78.9% and 76.3% mAP on PASCAL VOC 2007 and 2012 respectively.
****

Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778
Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers---8x deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.
****

You Only Look Once: Unified, Real-Time Object Detection
Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788
We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.
****

LocNet: Improving Localization Accuracy for Object Detection
Spyros Gidaris, Nikos Komodakis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 789-798
We propose a novel object localization methodology with the purpose of boosting the localization accuracy of state-of-the-art object detection systems. Our model, given a search region, aims at returning the bounding box of an object of interest inside this region. To accomplish its goal, it relies on assigning conditional probabilities to each row and column of this region, where these probabilities provide useful information regarding the location of the boundaries of the object inside the search region and allow the accurate inference of the object bounding box under a simple probabilistic framework. For implementing our localization model, we make use of a convolutional neural network architecture that is properly adapted for this task, called LocNet. We show experimentally that LocN

et achieves a very significant improvement on the mAP for high IoU thresholds on PASCAL VOC2007 test set and that it can be very easily coupled with recent state-of-the-art object detection systems, helping them to boost their performance. Finally, we demonstrate that our detection approach can achieve high detection accuracy even when it is given as input a set of sliding windows, thus proving that it is independent of box proposal methods.
****

Sketch Me That Shoe
Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, Chen-Change Loy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 799-807
We investigate the problem of fine-grained sketch-based image retrieval (SBIR), where free-hand human sketches are used as queries to perform instance-level retrieval of images. This is an extremely challenging task because (i) visual comparisons not only need to be fine-grained but also executed cross-domain, (ii) free-hand (finger) sketches are highly abstract, making fine-grained matching harder, and most importantly (iii) annotated cross-domain sketch-photo datasets required for training are scarce, challenging many state-of-the-art machine learning techniques. In this paper, for the first time, we address all these challenges, providing a step towards the capabilities that would underpin a commercial sketch-based image retrieval application. We introduce a new database of 1,432 sketch-photo pairs from two categories with 32,000 fine-grained triplet ranking annotations. We then develop a deep triplet-ranking model for instance-level SBIR with a novel data augmentation and staged pre-training strategy to alleviate the issue of insufficient fine-grained training data. Extensive experiments are carried out to contribute a variety of insights into the challenges of data sufficiency and over-fitting avoidance when training deep networks for fine-grained cross-domain ranking tasks.
****

Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images
Shuran Song, Jianxiong Xiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 808-816
We focus on the task of amodal 3D object detection in RGB-D images, which aims to produce a 3D bounding box of an object in metric form at its full extent. We introduce Deep Sliding Shapes, a 3D ConvNet formulation that takes a 3D volumetric scene from a RGB-D image as input and outputs 3D object bounding boxes. In our approach, we propose the first 3D Region Proposal Network (RPN) to learn objectness from geometric shapes and the first joint Object Recognition Network (ORN) to extract geometric features in 3D and color features in 2D. In particular, we handle objects of various sizes by training an amodal RPN at two different scales and an ORN to regress 3D bounding boxes. Experiments show that our algorithm outperforms the state-of-the-art by 13.8 in mAP and is 200x faster than the original Sliding Shapes.
****

Object Detection From Video Tubelets With Convolutional Neural Networks
Kai Kang, Wanli Ouyang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 817-825
Deep Convolution Neural Networks (CNNs) have shown impressive performance in various vision tasks such as image classification, object detection and semantic segmentation. For object detection, particularly in still images, the performance has been significantly increased last year thanks to powerful deep networks (e.g. GoogleNet) and detection frameworks (e.g. Regions with CNN features (R-CNN)). The lately introduced ImageNet task on object detection from video (VID) brings the object detection task into video domain, in which objects' locations at each frame are required to be annotated with bounding boxes. In this work, we introduce a complete framework for the VID task based on still-image object detection and general object tracking. Their relations and contributions in the VID task are thoroughly studied and evaluated. In addition, a temporal convolution network is proposed to incorporate temporal information to regularize the detection results and shows its effectiveness for the task.

****
Learning With Side Information Through Modality Hallucination
Judy Hoffman, Saurabh Gupta, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 826-834
We present a modality hallucination architecture for training an RGB object detection model which incorporates depth side information at training time. Our convolutional hallucination network learns a new and complementary RGB image representation which is taught to mimic convolutional mid-level features from a depth network. At test time images are processed jointly through the RGB and hallucination networks to produce improved detection performance. Thus, our method transfers information commonly extracted from depth training data to a network which can extract that information from the RGB counterpart. We present results on the standard NYUDv2 dataset and report improvement on the RGB detection task.
****
Object-Proposal Evaluation Protocol is 'Gameable'
Neelima Chavali, Harsh Agrawal, Aroma Mahendru, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 835-844
Object proposals have quickly become the de-facto pre-processing step in a number of vision pipelines (for object detection, object discovery, and other tasks). Their performance is usually evaluated on partially annotated datasets. In this paper, we argue that the choice of using a partially annotated dataset for evaluation of object proposals is problematic -- as we demonstrate via a thought experiment, the evaluation protocol is 'gameable', in the sense that progress under this protocol does not necessarily correspond to a "better" category independent object proposal algorithm. To alleviate this problem, we: (1) Introduce a nearly-fully annotated version of PASCAL VOC dataset, which serves as a test-bed to check if object proposal techniques are overfitting to a particular list of categories. (2) Perform an exhaustive evaluation of object proposal methods on our introduced nearly-fully annotated PASCAL dataset and perform cross-dataset generalization experiments; and (3) Introduce a diagnostic experiment to detect the bias capacity in an object proposal algorithm. This tool circumvents the need to collect a densely annotated dataset, which can be expensive and cumbersome to collect. Finally, we have released an easy-to-use toolbox which combines various publicly available implementations of object proposal algorithms which standardizes the proposal generation and evaluation so that new methods can be added and evaluated on different datasets. We hope that the results presented in the paper will motivate the community to test the category independence of various object proposal methods by carefully choosing the evaluation protocol.
****
HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection
Tao Kong, Anbang Yao, Yurong Chen, Fuchun Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 845-853
Almost all of the current top-performing object detection networks employ region proposals to guide the search for object instances. State-of-the-art region proposal methods usually need several thousand proposals to get high recall, thus hurting the detection efficiency. Although the latest Region Proposal Network method gets promising detection accuracy with several hundred proposals, it still struggles in small-size object detection and precise localization (e.g., large IoU thresholds), mainly due to the coarseness of its feature maps. In this paper, we present a deep hierarchical network, namely HyperNet, for handling region proposal generation and object detection jointly. Our HyperNet is primarily based on an elaborately designed Hyper Feature which aggregates hierarchical feature maps first and then compresses them into a uniform space. The Hyper Features well incorporate deep but highly semantic, intermediate but really complementary, and shallow but naturally high-resolution features of the image, thus enabling us to construct HyperNet by sharing them both in generating proposals and detecting objects via an end-to-end joint training strategy. For the deep VGG16 model, our method achieves completely leading recall and state-of-the-art object detection accuracy on PASCAL VOC 2007 and 2012 using only 100 proposals per image. It run

s with a speed of 5 fps (including all steps) on a GPU, thus having the potentia
l for real-time processing.
****

We Don't Need No Bounding-Boxes: Training Object Class Detectors Using Only Huma
n Verification

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, Vittorio Ferrari; Proc
eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
, 2016, pp. 854-863

Training object class detectors typically requires a large set of images in whic
h objects are annotated by bounding-boxes. However, manually drawing bounding-bo
xes is very time consuming. We propose a new scheme for training object detector
s which only requires annotators to verify bounding-boxes produced automatically
 by the learning algorithm. Our scheme iterates between re-training the detector
, re-localizing objects in the training images, and human verification. We use t
he verification signal both to improve re-training and to reduce the search spac
e for re-localisation, which makes these steps different to what is normally don
e in a weakly supervised setting. Extensive experiments on PASCAL VOC 2007 show
that (1) using human verification to update detectors and reduce the search spac
e leads to the rapid production of high-quality bounding-box annotations; (2) ou
r scheme delivers detectors performing almost as good as those trained in a full
y supervised setting, without ever drawing any bounding-box; (3) as the verifica
tion task is very quick, our scheme substantially reduces total annotation time
by a factor 6x-9x.
****

Factors in Finetuning Deep Model for Object Detection With Long-Tail Distributio
n

Wanli Ouyang, Xiaogang Wang, Cong Zhang, Xiaokang Yang; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 864-873

Finetuning from a pretrained deep model is found to yield state-of-the-art perfo
rmance for many vision tasks. This paper investigates many factors that influenc
e the performance in finetuning for object detection. There is a long-tailed d
istribution of sample numbers for classes in object detection. Our analysis and
empirical results show that classes with more samples have higher impact on the
feature learning. And it is better to make the sample number more uniform across
 classes. Generic object detection can be considered as multiple equally importa
nt tasks. Detection of each class is a task. These classes/tasks have their indi
viduality in discriminative visual appearance representation. Taking this indivi
duality into account, we cluster objects into visually similar class groups and
learn deep representations for these groups separately. A hierarchical feature l
earning scheme is proposed. In this scheme, the knowledge from the group with l
arge number of classes is transferred for learning features in its sub-groups. F
inetuned on the GoogLeNet model, experimental results show 4.7% absolute mAP imp
rovement of our approach on the ImageNet object detection dataset without increa
sing much computational cost at the testing stage.
****

Information-Driven Adaptive Structured-Light Scanners

Guy Rosman, Daniela Rus, John W. Fisher III; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 874-883

Sensor planning and active sensing, long studied in robotics, adapt sensor posit
ioning and operation mode in order to maximize information gain. While these con
cepts are often used to reason about 3D sensors, these are usually treated as a
predefined, black-box, component. In this paper we show how the same principles
can be used as part of the 3D sensor. We describe the relevant generative model
 for structured-light 3D scanning and show how adaptive pattern selection can ma
ximize information gain in an open-loop with-feedback manner. We then demonstrat
e how different choices of relevant variable sets (corresponding to the subprobl
ems of locatization and mapping) lead to different criteria for pattern selectio
n and can be computed in an online fashion. We show results for both subproblems
 with several pattern dictionary choices and demonstrate their usefulness for po
se estimation and depth acquisition.

****
Simultaneous Optical Flow and Intensity Estimation From an Event Camera
Patrick Bardow, Andrew J. Davison, Stefan Leutenegger; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 884-892
Event cameras are bio-inspired vision sensors which mimic retinas to measure per -pixel intensity change rather than outputting an actual intensity image. This p roposed paradigm shift away from traditional frame cameras offers significant po tential advantages: namely avoiding high data rates, dynamic range limitations a nd motion blur. Unfortunately, however, established computer vision algorithms m ay not at all be applied directly to event cameras. Methods proposed so far to reconstruct images, estimate optical flow, track a camera and reconstruct a scen e come with severe restrictions on the environment or on the motion of the camer a, e.g. allowing only rotation. Here, we propose, to the best of our knowledge, the first algorithm to simultaneously recover the motion field and brightness im age, while the camera undergoes a generic motion through any scene. Our approach employs minimisation of a cost function that contains the asynchronous event da ta as well as spatial and temporal regularisation within a sliding window time i nterval. Our implementation relies on GPU-based optimisation and runs in near re al-time. In a series of examples, we demonstrate the successful operation of our framework, including in situations where conventional cameras heavily suffer fr om dynamic range limitations or motion blur.
****
Macroscopic Interferometry: Rethinking Depth Estimation With Frequency-Domain Ti me-Of-Flight
Achuta Kadambi, Jamie Schiel, Ramesh Raskar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 893-902
A form of meter-scale, macroscopic interferometry is proposed using conventional time-of-flight (ToF) sensors. Today, ToF sensors use phase-based sampling, wher e the phase delay between emitted and received, high-frequency signals encodes d istance. This paper examines an alternative ToF architecture, inspired by micron -scale, microscopic interferometry, that relies only on frequency sampling: we r efer to our proposed macroscopic technique as Frequency-Domain Time of Flight (F D-ToF). The proposed architecture offers several benefits over existing phase To F systems, such as robustness to phase wrapping and implicit resolution of multi -path interference, all while capturing the same number of subframes. A prototyp e camera is constructed to demonstrate macroscopic interferometry at meter scale .
****
ASP Vision: Optically Computing the First Layer of Convolutional Neural Networks Using Angle Sensitive Pixels
Huaijin G. Chen, Suren Jayasuriya, Jiyue Yang, Judy Stephen, Sriram Sivaramakris hnan, Ashok Veeraraghavan, Alyosha Molnar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 903-912
Deep learning using convolutional neural networks (CNNs) is quickly becoming the state-of-the-art for challenging computer vision applications. However, deep le arning's power consumption and bandwidth requirements currently limit its applic ation in embedded and mobile systems with tight energy budgets. In this paper, w e explore the energy savings of optically computing the first layer of CNNs. To do so, we utilize bio-inspired Angle Sensitive Pixels (ASPs), custom CMOS diffra ctive image sensors which act similar to Gabor filter banks in the V1 layer of t he human visual cortex. ASPs replace both image sensing and the first layer of a conventional CNN by directly performing optical edge filtering, saving sensing energy, data bandwidth, and CNN FLOPS to compute. Our experimental results (both on synthetic data and a hardware prototype) for a variety of vision tasks such as digit recognition, object recognition, and face identification demonstrate 97 % reduction in image sensor power consumption and 90% reduction in data bandwidt h from sensor to CPU, while achieving similar performance compared to traditiona l deep learning pipelines.
****
Computational Imaging for VLBI Image Reconstruction

Katherine L. Bouman, Michael D. Johnson, Daniel Zoran, Vincent L. Fish, Sheperd S. Doeleman, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 913-922

Very long baseline interferometry (VLBI) is a technique for imaging celestial radio emissions by simultaneously observing a source from telescopes distributed across Earth. The challenges in reconstructing images from fine angular resolution VLBI data are immense. The data is extremely sparse and noisy, thus requiring statistical image models such as those designed in the computer vision community. In this paper we present a novel Bayesian approach for VLBI image reconstruction. While other methods often require careful tuning and parameter selection for different types of data, our method (CHIRP) produces good results under different settings such as low SNR or extended emission. The success of our method is demonstrated on realistic synthetic experiments as well as publicly available real data. We present this problem in a way that is accessible to members of the community, and provide a dataset website (vlbiimaging.csail.mit.edu) that facilitates controlled comparisons across algorithms.

****

You Lead, We Exceed: Labor-Free Video Concept Learning by Jointly Exploiting Web Videos and Images

Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 923-932

Video concept learning often requires a large set of training samples. In practice, however, acquiring noise-free training labels with sufficient positive examples is very expensive. A plausible solution for training data collection is by sampling from the vast quantities of images and videos on the Web. Such a solution is motivated by the assumption that the retrieved images or videos are highly correlated with the query. Still, a number of challenges remain. First, Web videos are often untrimmed. Thus, only parts of the videos are relevant to the query. Second, the retrieved Web images are always highly relevant to the issued query. However, thoughtlessly utilizing the images in the video domain may even hurt the performance due to the well known semantic drift and domain gap problems. As a result, a valid question is how Web images and videos interact for video concept learning. In this paper, we propose a Lead--Exceed Neural Network (LENN), which reinforces the training on Web images and videos in a curriculum manner. Specifically, the training proceeds by inputting frames of Web videos to obtain a network. The Web images are then filtered by the learnt network and the selected images are additionally fed into the network to enhance the architecture and further trim the videos. In addition, Long Short-Term Memory (LSTM) can be applied on the trimmed videos to explore temporal information. Encouraging results are reported on UCF101, TRECVID 2013 and 2014 MEDTest in the context of both action recognition and event detection. Without using human annotated exemplars, our proposed LENN can achieve 74.4% accuracy on UCF101 dataset.

****

Track and Segment: An Iterative Unsupervised Approach for Video Object Proposals

Fanyi Xiao, Yong Jae Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 933-942

We present an unsupervised approach that generates a diverse, ranked set of bounding box and segmentation video object proposals---spatio-temporal tubes that localize the foreground objects---in an unannotated video. In contrast to previous unsupervised methods that either track regions initialized in an arbitrary frame or train a fixed model over a cluster of regions, we instead discover a set of easy-to-group instances of an object and then iteratively update its appearance model to gradually detect harder instances in temporally-adjacent frames. Our method first generates a set of spatio-temporal bounding box proposals, and then refines them to obtain pixel-wise segmentation proposals. Through extensive experiments, we demonstrate state-of-the-art segmentation results on the SegTrack v2 dataset, and bounding box tracking results that perform competitively to state-of-the-art supervised tracking methods.

****

Beyond Local Search: Tracking Objects Everywhere With Instance-Specific Proposal

s
Gao Zhu, Fatih Porikli, Hongdong Li; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2016, pp. 943-951
Most tracking-by-detection methods employ a local search window around the predi cted object location in the current frame assuming the previous location is accu rate, the trajectory is smooth, and the computational capacity permits a search radius that can accommodate the maximum speed yet small enough to reduce mismatc hes. These, however, may not be valid always, in particular for fast and irregul arly moving objects. Here, we present an object tracker that is not limited to a local search window and has ability to probe efficiently the entire frame. Our method generates a small number of "high-quality" proposals by a novel instance-specific objectness measure and evaluates them against the object model that can be adopted from an existing tracking-by-detection approach as a core tracker. D uring the tracking process, we update the object model concentrating on hard fal se-positives supplied by the proposals, which help suppressing distractors cause d by difficult background clutters, and learn how to re-rank proposals according to the object model. Since we reduce significantly the number of hypotheses the core tracker evaluates, we can use richer object descriptors and stronger detec tor. Our method outperforms most recent state-of-the-art trackers on popular tr acking benchmarks, and provides improved robustness for fast moving objects as w ell as for ultra low-frame-rate videos.
****
Groupwise Tracking of Crowded Similar-Appearance Targets From Low-Continuity Ima ge Sequences
Hongkai Yu, Youjie Zhou, Jeff Simmons, Craig P. Przybyla, Yuewei Lin, Xiaochuan Fan, Yang Mi, Song Wang; Proceedings of the IEEE Conference on Computer Vision a nd Pattern Recognition (CVPR), 2016, pp. 952-960
Automatic tracking of large-scale crowded targets are of particular importance i n many applications, such as crowded people/vehicle tracking in video surveillan ce, fiber tracking in materials science, and cell tracking in biomedical imaging . This problem becomes very challenging when the targets show similar appearance and the inter-slice/inter-frame continuity is low due to sparse sampling, camer a motion and target occlusion. The main challenge comes from the step of associa tion which aims at matching the predictions and the observations of the multiple targets. In this paper we propose a new groupwise method to explore the target group information and employ the within-group correlations for association and t racking. In particular, the within-group association is modeled by a nonrigid 2D Thin-Plate transform and a sequence of group shrinking, group growing and group merging operations are then developed to refine the composition of each group. We apply the propose method to track large-scale fibers from the microscopy mate rial images and compare its performance against several other multi-target track ing methods. We also apply the proposed method to track crowded people from vide os with poor inter-frame continuity.
****
Social LSTM: Human Trajectory Prediction in Crowded Spaces
Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 961-971
Humans navigate complex crowded environments based on social conventions: they r espect personal space, yielding right-of-way and avoid collisions. In our work, we propose a data-driven approach to learn these human-human interactions for pr edicting their future trajectories. This is in contrast to traditional approache s which use hand-crafted functions such as Social forces. We present a new Long Short-Term Memory (LSTM) model which jointly reasons across multiple individual s in a scene. Different from the conventional LSTM, we share the information bet ween multiple LSTMs through a new pooling layer. This layer pools the hidden rep resentation from LSTMs corresponding to neighboring trajectories to capture inte ractions within this neighborhood. We demonstrate the performance of our method on several public datasets. Our model outperforms previous forecasting methods b y more than 42% . We also analyze the trajectories predicted by our model to dem

onstrate social behaviours such as collision avoidance and group movement, learned by our model.
****
What Players Do With the Ball: A Physically Constrained Interaction Modeling
Andrii Maksai, Xinchao Wang, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 972-981
Tracking the ball is critical for video-based analysis of team sports. However, it is difficult, especially in low-resolution images, due to the small size of the ball, its speed that creates motion blur, and its often being occluded by players. In this paper, we propose a generic and principled approach to modeling the interaction between the ball and the players while also imposing appropriate physical constraints on the ball's trajectory. We show that our approach, formulated in terms of a Mixed Integer Program, is more robust and more accurate than several state-of-the-art approaches on real-life volleyball, basketball, and soccer sequences.
****
Highlight Detection With Pairwise Deep Ranking for First-Person Video Summarization
Ting Yao, Tao Mei, Yong Rui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 982-990
The emergence of wearable devices such as portable cameras and smart glasses makes it possible to record life logging first-person videos. Browsing such long unstructured videos is time-consuming and tedious. This paper studies the discovery of moments of user's major or special interest (i.e., highlights) in a video, for generating the summarization of first-person videos. Specifically, we propose a novel pairwise deep ranking model that employs deep learning techniques to learn the relationship between highlight and non-highlight video segments. A two-stream network structure by representing video segments from complementary information on appearance of video frames and temporal dynamics across frames is developed for video highlight detection. Given a long personal video, equipped with the highlight detection model, a highlight score is assigned to each segment. The obtained highlight segments are applied for summarization in two ways: video timelapse and video skimming. The former plays the highlight (non-highlight) segments at low (high) speed rates, while the latter assembles the sequence of segments with the highest scores. On 100 hours of first-person videos for 15 unique sports categories, our highlight detection achieves the improvement over the state-of-the-art RankSVM method by 10.5% in terms of accuracy. Moreover, our approaches produce video summary with better quality by a user study from 35 human subjects.
****
Direct Prediction of 3D Body Poses From Motion Compensated Sequences
Bugra Tekin, Artem Rozantsev, Vincent Lepetit, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 991-1000
We propose an efficient approach to exploiting motion information from consecutive frames of a video sequence to recover the 3D pose of people. Previous approaches typically compute candidate poses in individual frames and then link them in a post-processing step to resolve ambiguities. By contrast, we directly regress from a spatio-temporal volume of bounding boxes to a 3D pose in the central frame. We further show that, for this approach to achieve its full potential, it is essential to compensate for the motion in consecutive frames so that the subject remains centered. This then allows us to effectively overcome ambiguities and improve upon the state-of-the-art by a large margin on the Human3.6m, HumanEva, and KTH Multiview Football 3D human pose estimation benchmarks.
****
Video2GIF: Automatic Generation of Animated GIFs From Video
Michael Gygli, Yale Song, Liangliang Cao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1001-1009
We introduce the novel problem of automatically generating animated GIFs from video. GIFs are short looping video with no sound, and a perfect combination betwe

en image and video that really capture our attention. GIFs tell a story, express emotion, turn events into humorous moments, and are the new wave of photojournalism. We pose the question: Can we automate the entirely manual and elaborate process of GIF creation by leveraging the plethora of user generated GIF content? We propose a Robust Deep RankNet that, given a video, generates a ranked list of its segments according to their suitability as GIF. We train our model to learn what visual content is often selected for GIFs by using over 100K user generated GIFs and their corresponding video sources. We effectively deal with the noisy web data by proposing a novel adaptive Huber loss in the ranking formulation. We show that our approach is robust to outliers and picks up several patterns that are frequently present in popular animated GIFs. On our new large-scale benchmark dataset, we show the advantage of our approach over several state-of-the-art methods.

****

NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010-1019

Recent approaches in depth-based human activity analysis achieved outstanding performance and proved the effectiveness of 3D representation for classification of action classes. Currently available depth-based and RGB+D-based action recognition benchmarks have a number of limitations, including the lack of training samples, distinct class labels, camera views and variety of subjects. In this paper we introduce a large-scale dataset for RGB+D human action recognition with more than 56 thousand video samples and 4 million frames, collected from 40 distinct subjects. Our dataset contains 60 different action classes including daily, mutual, and health-related actions. In addition, we propose a new recurrent neural network structure to model the long-term temporal correlation of the features for each body part, and utilize them for better action classification. Experimental results show the advantages of applying deep learning methods over state-of-the-art hand-crafted features on the suggested cross-subject and cross-view evaluation criteria for our dataset. The introduction of this large scale dataset will enable the community to apply, develop and adapt various data-hungry learning techniques for the task of depth-based and RGB+D-based human activity analysis.

****

Progressively Parsing Interactional Objects for Fine Grained Action Detection

Bingbing Ni, Xiaokang Yang, Shenghua Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1020-1028

Fine grained video action analysis often requires reliable detection and tracking of various interacting objects and human body parts, denoted as interactional object parsing. However, most of the previous methods based on either independent or joint object detection might suffer from high model complexity and challenging image content, e.g., illumination/pose/appearance/scale variation, motion, occlusion etc. In this work, we propose an end-to-end system based on recursive neural network to perform frame by frame interactional object parsing, which can alleviate the difficulty through a incremental manner. Our key innovation is that: instead of jointly outputting all object detections at once, for each frame, we use a set of long-short term memory (LSTM) nodes to incrementally refine the detections. After passing each LSTM node, more object detections are consolidated and thus more contextual information could be utilized to determine more difficult object detections. Extensive experiments on two benchmark fine grained activity datasets demonstrate that our proposed algorithm achieves better interacting object detection performance, which in turn boosts the action recognition performance over the state-of-the-art.

****

Hierarchical Recurrent Neural Encoder for Video Representation With Application to Captioning

Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, Yueting Zhuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1029-1038

Recently, deep learning approach, especially deep Convolutional Neural Networks

(ConvNets), have achieved overwhelming accuracy with fast processing speed for image classification. Incorporating temporal structure with deep ConvNets for video representation becomes a fundamental problem for video content analysis. In this paper, we propose a new approach, namely Hierarchical Recurrent Neural Encoder (HRNE), to exploit temporal information of videos. Compared to recent video representation inference approaches, this paper makes the following three contributions. First, our HRNE is able to efficiently exploit video temporal structure in a longer range by reducing the length of input information flow, and compositing multiple consecutive inputs at a higher level. Second, computation operations are significantly lessened while attaining more non-linearity. Third, HRNE is able to uncover temporal transitions between frame chunks with different granularities, i.e. it can model the temporal transitions between frames as well as the transitions between segments. We apply the new method to video captioning where temporal information plays a crucial role. Experiments demonstrate that our method outperforms the state-of-the-art on video captioning benchmarks.

****

From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection

Jingjing Meng, Hongxing Wang, Junsong Yuan, Yap-Peng Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1039-1048

We propose to summarize a video into a few key objects by selecting representative object proposals generated from video frames. This representative selection problem is formulated as a sparse dictionary selection problem, i.e., choosing a few representatives object proposals to reconstruct the whole proposal pool. Compared with existing sparse dictionary selection based representative selection methods, our new formulation can incorporate object proposal priors and locality prior in the feature space when selecting representatives. Consequently it can better locate key objects and suppress outlier proposals. We convert the optimization problem into a proximal gradient problem and solve it by the fast iterative shrinkage thresholding algorithm (FISTA). Experiments on synthetic data and real benchmark datasets show promising results of our key object summarization approach in video content mining and search. Comparisons with existing representative selection approaches such as K-mediod, sparse dictionary selection and density based selection validate that our formulation can better capture the key video objects despite appearance variations, cluttered backgrounds and camera motions.

****

Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs

Zheng Shou, Dongang Wang, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1049-1058

We address temporal action localization in untrimmed long videos. This is important because videos in real applications are usually unconstrained and contain multiple action instances plus video content of background scenes or other activities. To address this challenging issue, we exploit the effectiveness of deep networks in temporal action localization via three segment-based 3D ConvNets: (1) a proposal network identifies candidate segments in a long video that may contain actions; (2) a classification network learns one-vs-all action classification model to serve as initialization for the localization network; and (3) a localization network fine-tunes the learned classification network to localize each action instance. We propose a novel loss function for the localization network to explicitly consider temporal overlap and achieve high temporal localization accuracy. In the end, only the proposal network and the localization network are used during prediction. On two large-scale benchmarks, our approach achieves significantly superior performances compared with other state-of-the-art systems: mAP increases from 1.7% to 7.4% on MEXaction2 and increases from 15.0% to 19.0% on THUMOS 2014.

****

Summary Transfer: Exemplar-Based Subset Selection for Video Summarization

Ke Zhang, Wei-Lun Chao, Fei Sha, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1059-1067

Video summarization has unprecedented importance to help us digest, browse, and search today's ever-growing video collections. We propose a novel subset selecti on technique that leverages supervision in the form of human-created summaries t o perform automatic keyframe-based video summarization. The main idea is to nonp arametrically transfer summary structures from annotated videos to unseen test v ideos. We show how to extend our method to exploit semantic side information abo ut the video's category/genre to guide the transfer process by those training vi deos semantically consistent with the test input. We also show how to generalize our method to subshot-based summarization, which not only reduces computational costs but also provides more flexible ways of defining visual similarity across subshots spanning several frames. We conduct extensive evaluation on several be nchmarks and demonstrate promising results, outperforming existing methods in se veral settings.
****

POD: Discovering Primary Objects in Videos Based on Evolutionary Refinement of O bject Recurrence, Background, and Primary Object Models
Yeong Jun Koh, Won-Dong Jang, Chang-Su Kim; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1068-1076
A primary object discovery (POD) algorithm for a video sequence is proposed in t his work, which is capable of discovering a primary object, as well as identifyi ng noisy frames that do not contain the object. First, we generate object propos als for each frame. Then, we bisect each proposal into foreground and background regions, and extract features from each region. By superposing the foreground a nd background features, we build the object recurrence model, the background mod el, and the primary object model. We develop an iterative scheme to refine each model evolutionary using the information in the other models. Finally, using the evolved primary object model, we select candidate proposals and locate the boun ding box of a primary object by merging the proposals selectively. Experimental results on a challenging dataset demonstrate that the proposed POD algorithm ext ract primary objects accurately and robustly.
****

What If We Do Not Have Multiple Videos of the Same Action? -- Video Action Local ization Using Web Images
Waqas Sultani, Mubarak Shah; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2016, pp. 1077-1085
This paper tackles the problem of spatio-temporal action localization in a video without assuming the availability of multiple videos or any prior annotations. Action is localized by employing images downloaded from internet using action na me. Given web images, we first mitigate image noise using random walk framework and evade distracting backgrounds within images using image action proposals. T hen, given a video, we generate multiple spatio-temporal action proposals. We s uppress camera and background generated proposals by exploiting optical flow gra dients within proposal. To obtain the most action representative proposal, we pr opose to reconstruct action proposals in the video by leveraging the action pro posal in images. Moreover, we preserve the temporal smoothness of the video by i ntroducing consensus regularization. Consensus regularization enforces consisten cy among coefficients vectors of multiple frames within proposal. %We reconstruc t video action proposals from image action proposals while enforcing consistency across coefficient vectors of multiple frames by consensus regularization. Fin ally, the video proposal that have the lowest reconstruction cost and is motion salient is considered as final action localization. Our extensive experiments o n trimmed as well as untrimmed datasets validate the effectiveness of proposed a pproach.
****

Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups From Static Images
Lu Zhang, Hayley Hung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1086-1095
In this paper, we present the first attempt to analyse differing levels of socia l involvement in free standing conversing groups (or the so-called F-formations)

from static images. In addition, we enrich state-of-the-art F-formation modelling by learning a frustum of attention that accounts for the spatial context. That is, F-formation configurations vary with respect to the arrangement of furniture and the non-uniform crowdedness in the space during mingling scenarios. The majority of prior works have considered the labelling of conversing group as an objective task, requiring only a single annotator. However, we show that by embracing the subjectivity of social involvement, we not only generate a richer model of the social interactions in a scene but also significantly improve F-formation detection. We carry out extensive experimental validation of our proposed approach by collecting a novel set of multi-annotator labels of involvement on the publicly available Idiap Poster Data; the only multi-annotator labelled database of free standing conversing groups that is currently available.
****

DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations
Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096-1104
Recent advances in clothes recognition have been driven by the construction of clothes datasets. Existing datasets are limited in the amount of annotations and are difficult to cope with the various challenges in real-world applications. In this work, we introduce DeepFashion, a large-scale clothes dataset with comprehensive annotations. It contains over 800,000 images, which are richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. Such rich annotations enable the development of powerful algorithms in clothes recognition and facilitating future researches. To demonstrate the advantages of DeepFashion, we propose a new deep model, namely FashionNet, which learns clothing features by jointly predicting clothing attributes and landmarks. The estimated landmarks are then employed to pool or gate the learned features. It is optimized in an iterative manner. Extensive experiments demonstrate the effectiveness of FashionNet and the usefulness of DeepFashion.
****

SketchNet: Sketch Classification With Web Images
Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, Xiaochun Cao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1105-1113
In this study, we present a weakly supervised approach that discovers the discriminative structures of sketch images, given pairs of sketch images and web images. In contrast to traditional approaches that use global appearance features or relay on keypoint features, our aim is to automatically learn the shared latent structures that exist between sketch images and real images, even when there are significant appearance differences across its relevant real images. To accomplish this, we propose a deep convolutional neural network, named SketchNet. We firstly develop a triplet composed of sketch, positive and negative real image as the input of our neural network. To discover the coherent visual structures between the sketch and its positive pairs, we introduce the softmax as the loss function. Then a ranking mechanism is introduced to make the positive pairs obtain a higher score comparing over negative ones to achieve robust representation. Finally, we formalize above-mentioned constrains into the unified objective function, and create an ensemble feature representation to describe the sketch images. Experiments on the TU-Berlin sketch benchmark demonstrate the effectiveness of our model and show that deep feature representation brings substantial improvements over other state-of-the-art methods on sketch classification.
****

Embedding Label Structures for Fine-Grained Feature Representation
Xiaofan Zhang, Feng Zhou, Yuanqing Lin, Shaoting Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1114-1123
Recent algorithms in convolutional neural networks (CNN) considerably advance th

e fine-grained image classification, which aims to differentiate the subtle diff
erences among subordinate classes. However, previous studies have rarely focused
 on learning a fined-grained and structured feature representation that is able
to locate relevant images at different levels of relevance, e.g., discovering ca
rs from the same make or the same model, both of which require high precision. I
n this paper, we propose two main contributions to tackle this problem. 1) A mul
ti-task learning framework is designed to effectively learn fine-grained feature
 representations by jointly optimizing both classification and similarity constr
aints. 2) To model the multi-level relevance, label structures such as hierarchy
 or shared attributes are seamlessly embedded into the framework by generalizing
 the triplet loss. Extensive and thorough experiments have been conducted on thr
ee fine-grained datasets, i.e., the Stanford car, the car-333, and the food data
sets, which contain either hierarchical labels or shared attributes. Our propose
d method has achieved very competitive performance, i.e., among state-of-the-art
 classification accuracy. More importantly, it significantly outperforms previou
s fine-grained feature representations for image retrieval at different levels o
f relevance
****

Fine-Grained Image Classification by Exploring Bipartite-Graph Labels
Feng Zhou, Yuanqing Lin; Proceedings of the IEEE Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2016, pp. 1124-1133
Given a food image, can a fine-grained object recognition engine tell "which res
taurant which dish" the food belongs to? Such ultra-fine grained image recogniti
on is the key for many applications like search by images, but it is very challe
nging because it needs to discern subtle difference between classes while dealin
g with the scarcity of training data. Fortunately, the ultra-fine granularity na
turally brings rich relationships among object classes. This paper proposes a no
vel approach to exploit the rich relationships through bipartite-graph labels (B
GL). We show how to model BGL in an overall convolutional neural networks and th
e resulting system can be optimized through back-propagation. We also show that
it is computationally efficient in inference thanks to the bipartite structure.
To facilitate the study, we construct a new food benchmark dataset, which consis
ts of 37,885 food images collected from 6 restaurants and totally 975 menus. Exp
erimental results on this new food and three other datasets demonstrate BGL adva
nces previous works in fine-grained object recognition. An online demo is availa
ble at http://www.f-zhou.com/fg_demo/.
****

Picking Deep Filter Responses for Fine-Grained Image Recognition
Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, Qi Tian; Proceedings of
 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp
. 1134-1142
Recognizing fine-grained sub-categories such as birds and dogs is extremely chal
lenging due to the highly localized and subtle differences in some specific part
s. Most previous works rely on object/part level annotations to build part-based
 representation, which is demanding in practical applications. This paper propos
es an automatic fine-grained recognition approach which is free of any object/pa
rt annotation at both training and testing stages. Our method explores a unified
 framework based on two steps of deep filter response picking. The first picking
 step is to find distinctive filters which respond to specific patterns signific
antly and consistently, and learn a set of part detectors via iteratively altern
ating between new positive sample mining and part model retraining. The second p
icking step is to pool deep filter responses via spatially weighted combination
of Fisher Vectors. We conditionally pick deep filter responses to encode them in
to the final representation, which considers the importance of filter responses
themselves. Integrating all these techniques produces a much more powerful frame
work, and experiments conducted on CUB-200-2011 and Stanford Dogs demonstrate th
e superiority of our proposed algorithm over the existing methods.
****

SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Reco
gnition

Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, Dimitris Metaxas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1143-1152

Most convolutional neural networks (CNNs) lack midlevel layers that model semantic parts of objects. This limits CNN-based methods from reaching their full potential in detecting and utilizing small semantic parts in recognition. Introducing such mid-level layers can facilitate the extraction of part-specific features which can be utilized for better recognition performance. This is particularly important in the domain of fine-grained recognition. In this paper, we propose a new CNN architecture that integrates semantic part detection and abstraction (SPDA-CNN) for fine-grained classification. The proposed network has two sub-networks: one for detection and one for recognition. The detection sub-network has a novel top-down proposal method to generate small semantic part candidates for detection. The classification sub-network introduces novel part layers that extract features from parts detected by the detection sub-network, and combine them for recognition. As a result, the proposed architecture provides an end-to-end network that performs detection, localization of multiple semantic parts, and whole object recognition within one framework that shares the computation of convolutional filters. Our method outperforms state-of-the-art methods with a large margin for small parts detection (e.g. our precision of 93.40% vs the best previous precision of 74.00% for detecting the head on CUB-2011). It also compares favorably to the existing state-of-the-art on fine-grained classification, e.g. it achieves 85.14% accuracy on CUB-2011.

****

Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning With Humans in the Loop

Yin Cui, Feng Zhou, Yuanqing Lin, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1153-1162

Existing fine-grained visual categorization methods often suffer from three challenges: lack of training data, large number of fine-grained categories, and high intra-class vs. low inter-class variance. In this work we propose a generic iterative framework for fine-grained categorization and dataset bootstrapping that handles these three challenges. Using deep metric learning with humans in the loop, we learn a low dimensional feature embedding with anchor points on manifolds for each category. These anchor points capture intra-class variances and remain discriminative between classes. In each round, images with high confidence scores from our model are sent to humans for labeling. By comparing with exemplar images, labelers mark each candidate image as either a "true positive" or a "false positive." True positives are added into our current dataset and false positives are regarded as "hard negatives" for our metric learning model. Then the model is re-trained with an expanded dataset and hard negatives for the next round. To demonstrate the effectiveness of the proposed framework, we bootstrap a fine-grained flower dataset with 620 categories from Instagram images. The proposed deep metric learning scheme is evaluated on both our dataset and the CUB-200-2001 Birds dataset. Experimental evaluations show significant performance gain using dataset bootstrapping and demonstrate state-of-the-art results achieved by the proposed deep metric learning methods.

****

Mining Discriminative Triplets of Patches for Fine-Grained Classification

Yaming Wang, Jonghyun Choi, Vlad Morariu, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1163-1172

Fine-grained classification involves distinguishing between similar sub-categories based on subtle differences in highly localized regions; therefore, accurate localization of discriminative regions remains a major challenge. We describe a patch-based framework to address this problem. We introduce triplets of patches with geometric constraints to improve the accuracy of patch localization, and automatically mine discriminative geometrically-constrained triplets for classification. The resulting approach only requires object bounding boxes. Its effectiveness is demonstrated using four publicly available fine-grained datasets, on wh

ich it outperforms or obtains comparable results to the state-of-the-art in clas
sification.
****

Part-Stacked CNN for Fine-Grained Visual Categorization
Shaoli Huang, Zhe Xu, Dacheng Tao, Ya Zhang; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1173-1182
In the context of fine-grained visual categorization, the ability to interpret m
odels as human-understandable visual manuals is sometimes as important as achiev
ing high classification accuracy. In this paper, we propose a novel Part-Stacked
 CNN architecture that explicitly explains the fine-grained recognition process
by modeling subtle differences from object parts. Based on manually-labeled stro
ng part annotations, the proposed architecture consists of a fully convolutional
 network to locate multiple object parts and a two-stream classification network
 that encodes object-level and part-level cues simultaneously. By adopting a set
 of sharing strategies between the computation of multiple object parts, the pro
posed architecture is very efficient running at 20 frames/sec during inference.
Experimental results on the CUB-200-2011 dataset reveal the effectiveness of the
 proposed architecture, from multiple perspectives of classification accuracy, m
odel interpretability, and efficiency. Being able to provide interpretable recog
nition results in realtime, the proposed method is believed to be effective in p
ractical applications.
****

Learning Compact Binary Descriptors With Unsupervised Deep Neural Networks
Kevin Lin, Jiwen Lu, Chu-Song Chen, Jie Zhou; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1183-1192
In this paper, we propose a new unsupervised deep learning approach called DeepB
it to learn compact binary descriptor for efficient visual object matching. Unli
ke most existing binary descriptors which were designed with random projections
or linear hash functions, we develop a deep neural network to learn binary descr
iptors in a unsupervised manner. We enforce three criterions on binary codes whi
ch are learned at the top layer of our network: 1) minimal loss quantization, 2)
 evenly distributed codes and 3) uncorrelated bits. Then, we learn the parameter
s of the networks with a back-propagation technique. Experimental results on thr
ee different visual analysis tasks including image matching, image retrieval, an
d object recognition clearly demonstrate the effectiveness of the proposed appro
ach.
****

Solving Small-Piece Jigsaw Puzzles by Growing Consensus
Kilho Son, daniel Moreno, James Hays, David B. Cooper; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1193-1201
In this paper, we present a novel computational puzzle solver for square-piece i
mage jigsaw puzzles with no prior information such as piece orientation, anchor
pieces or resulting dimension of the puzzle. By "piece" we mean a square dxd blo
ck of pixels, where we investigate pieces as small as 7x7 pixels. To reconstruct
 such challenging puzzles, we aim to search for piece configurations which maxim
ize the size of consensus (i.e. grid or loop) configurations which represent a g
eometric consensus or agreement among pieces. Pieces are considered for addition
 to the existing assemblies if these pieces increase the size of the consensus c
onfigurations. In contrast to previous puzzle solvers which goal for assemblies
maximizing compatibility measures between all pairs of pieces and thus depend he
avily on the pairwise compatibility measure used, our new approach reduces the d
ependency on the pairwise compatibility measures which become increasingly uninf
ormative at small scales and instead exploits geometric agreement among pieces.
Our contribution also includes an improved pairwise compatibility measure which
exploits directional derivative information along adjoining boundaries of the pi
eces. For the challenging unknown orientation piece puzzles where the size of pi
eces is small, we reduce assembly error by up to 75% compared with previous algo
rithms for standard datasets.
****

Pairwise Matching Through Max-Weight Bipartite Belief Propagation

Zhen Zhang, Qinfeng Shi, Julian McAuley, Wei Wei, Yanning Zhang, Anton van den H
engel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2016, pp. 1202-1210
Feature matching is a key problem in computer vision and pattern recognition. On
e way to encode the essential interdependence between potential feature matches
is to cast the problem as inference in a graphical model, though recently altern
atives such as spectral methods, or approaches based on the convex-concave proce
dure have achieved the state-of-the-art. Here we revisit the use of graphical mo
dels for feature matching, and propose a belief propagation scheme which exhibit
s the following advantages: (1) we explicitly enforce one-to-one matching constr
aints; (2) we offer a tighter relaxation of the original cost function than prev
ious graphical-model-based approaches; and (3) our sub-problems decompose into m
ax-weight bipartite matching, which can be solved efficiently, leading to orders
-of-magnitude reductions in execution time. Experimental results show that the p
roposed algorithm produces results superior to those of the current state-of-the
-art.
****

Structured Feature Similarity With Explicit Feature Map
Takumi Kobayashi; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2016, pp. 1211-1219
Feature matching is a fundamental process in a variety of computer vision tasks.
 Beyond the standard L2 metric, various methods to measure similarity between fe
atures have been proposed mainly on the assumption that the features are defined
 in a histogram form. On the other hand, in a field of image quality assessment,
 SSIM produces effective similarity between images, taking the place of L2 metri
c. In this paper, we propose a feature similarity measurement method based on th
e SSIM. Unlike the previous methods, the proposed method is built on not a histo
gram form but a tensor structure of a feature array extracted such as on spatial
 grids, in order to construct effective SSIM-based similarity measure of high ro
bustness which is a key requirement in feature matching. In addition, we provide
 the explicit feature map such that the proposed similarity metric is embedded a
s a dot product. It contributes to significant speedup in similarity measurement
 as well as to feature transformation toward an effective vector form to which l
inear classifiers are directly applicable. In the experiments on various tasks,
the proposed method exhibits favorable performance in both feature matching and
classification.
****

Temporal Epipolar Regions
Mor Dar, Yael Moses; Proceedings of the IEEE Conference on Computer Vision and P
attern Recognition (CVPR), 2016, pp. 1220-1228
Dynamic events are often photographed by a number of people from different viewp
oints at different times, resulting in an unconstrained set of images. Finding t
he corresponding moving features in each of the images allows us to extract info
rmation about objects of interest in the scene. Computing correspondence of movi
ng features in such a set of images is considerably more challenging than comput
ing correspondence in video due to possible significant differences in viewpoint
s and inconsistent timing between image captures. The prediction methods used in
 video for improving robustness and efficiency are not applicable to a set of st
ill images. In this paper we propose a novel method to predict locations of an a
pproximately linear moving feature point, given a small subset of correspondence
s and the temporal order of image captures. Our method extends the use of epipol
ar geometry to divide images into valid and invalid regions, termed Temporal Epi
polar Regions (TERs). We formally prove that the location of a feature in a new
image is restricted to valid TERs. We demonstrate the effectiveness of our metho
d in reducing the search space for correspondence on both synthetic and challeng
ing real world data, and show the improved matching.
****

Recurrent Attention Models for Depth-Based Person Identification
Albert Haque, Alexandre Alahi, Li Fei-Fei; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1229-1238

We present an attention-based model that reasons on human body shape and motion dynamics to identify individuals in the absence of RGB information, hence in the dark. Our approach leverages unique 4D spatio-temporal signatures to address the identification problem across days. Formulated as a reinforcement learning task, our model is based on a combination of convolutional and recurrent neural networks with the goal of identifying small, discriminative regions indicative of human identity. We demonstrate that our model produces state-of-the-art results on several published datasets given only depth images. We further study the robustness of our model towards viewpoint, appearance, and volumetric changes. Finally, we share insights gleaned from interpretable 2D, 3D, and 4D visualizations of our model's spatio-temporal attention.
****

Learning a Discriminative Null Space for Person Re-Identification
Li Zhang, Tao Xiang, Shaogang Gong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1239-1248
Most existing person re-identification (re-id) methods focus on learning the optimal distance metrics across camera views. Typically a person's appearance is represented using features of thousands of dimensions, whilst only hundreds of training samples are available due to the difficulties in collecting matched training images. With the number of training samples much smaller than the feature dimension, the existing methods thus face the classic small sample size (SSS) problem and have to resort to dimensionality reduction techniques and/or matrix regularisation, which lead to loss of discriminative power. In this work, we propose to overcome the SSS problem in re-id distance metric learning by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-class separation simultaneously. Importantly, it has a fixed dimension, a closed-form solution and is very efficient to compute. Extensive experiments carried out on five person re-identification benchmarks including VIPeR, PRID2011, CUHK01, CUHK03 and Market1501 show that such a simple approach beats the state-of-the-art alternatives, often by a big margin.
****

Learning Deep Feature Representations With Domain Guided Dropout for Person Re-Identification
Tong Xiao, Hongsheng Li, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1249-1258
Learning generic and robust feature representations with data from multiple domains for the same problem is of great value, especially for the problems that have multiple datasets but none of them are large enough to provide abundant data variations. In this work, we present a pipeline for learning deep feature representations from multiple domains with Convolutional Neural Networks (CNNs). When training a CNN with data from all the domains, some neurons learn representations shared across several domains, while some others are effective only for a specific one. Based on this important observation, we propose a Domain Guided Dropout algorithm to improve the feature learning procedure. Experiments show the effectiveness of our pipeline and the proposed algorithm. Our methods on the person re-identification problem outperform state-of-the-art methods on multiple datasets by large margins.
****

How Far Are We From Solving Pedestrian Detection?
Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1259-1267
Encouraged by the recent progress in pedestrian detection, we investigate the gap between current state-of-the-art methods and the "perfect single frame detector". We enable our analysis by creating a human baseline for pedestrian detection (over the Caltech dataset), and by manually clustering the recurrent errors of a top detector. Our results characterise both localisation and background-versus-foreground errors. To address localisation errors we study the impact of train

ing annotation noise on the detector performance, and show that we can improve e ven with a small portion of sanitised training data. To address background/foreg round discrimination, we study convnets for pedestrian detection, and discuss wh ich factors affect their performance. Other than our in-depth analysis, we repo rt top performance on the Caltech dataset, and provide a new sanitised set of tr aining and test annotations.
****

Similarity Learning With Spatial Constraints for Person Re-Identification
Dapeng Chen, Zejian Yuan, Badong Chen, Nanning Zheng; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1268-1277
Pose variation remains one of the major factors that adversely affect the accura cy of person re-identification. Such variation is not arbitrary as body parts (e .g. head, torso, legs) have relative stable spatial distribution. Breaking down the variability of global appearance regarding the spatial distribution potentia lly benefits the person matching. We therefore learn a novel similarity function , which consists of multiple sub-similarity measurements with each taking in cha rge of a subregion. In particular, we take advantage of the recently proposed po lynomial feature map to describe the matching within each subregion, and inject all the feature maps into a unified framework. The framework not only outputs s imilarity measurements for different regions, but also makes a better consistenc y among them. Our framework can collaborate local similarities as well as global similarity to exploit their complementary strength. It is flexible to incorpora te multiple visual cues to further elevate the performance. In experiments, we a nalyze the effectiveness of the major components. The results on four datasets s how significant and consistent improvements over the state-of-the-art methods.
****

Sample-Specific SVM Learning for Person Re-Identification
Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, Xiang Ruan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 12 78-1287
Person re-identification addresses the problem of matching people across disjoin t camera views and extensive efforts have been made to seek either the robust fe ature representation or the discriminative matching metrics. However, most exist ing approaches focus on learning a fixed distance metric for all instance pairs, while ignoring the individuality of each person. In this paper, we formulate th e person re-identification problem as an imbalanced classification problem and l earn a classifier specifically for each pedestrian such that the matching model is highly tuned to the individual's appearance. To establish correspondence betw een feature space and classifier space, we propose a Least Square Semi-Coupled D ictionary Learning (LSSCDL) algorithm to learn a pair of dictionaries and a mapp ing function efficiently. Extensive experiments on a series of challenging datab ases demonstrate that the proposed algorithm performs favorably against the stat e-of-the-art approaches, especially on the rank-1 recognition rate.
****

Joint Learning of Single-Image and Cross-Image Representations for Person Re-Ide ntification
Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, Lei Zhang; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1 288-1296
Person re-identification has been usually solved as either the matching of singl e-image representation (SIR) or the classification of cross-image representation (CIR). In this work, we exploit the connection between these two categories of methods, and propose a joint learning framework to unify SIR and CIR using convo lutional neural network (CNN). Specifically, our deep architecture contains one shared sub-network together with two sub-networks that extract the SIRs of given images and the CIRs of given image pairs, respectively. The SIR sub-network is required to be computed once for each image (in both the probe and gallery sets) , and the depth of the CIR sub-network is required to be minimal to reduce compu tational burden. Therefore, the two types of representation can be jointly optim ized for pursuing better matching accuracy with moderate computational cost. Fur

thermore, the representations learned with pairwise comparison and triplet compa
rison objectives can be combined to improve matching performance. Experiments on
 the CUHK03, CUHK01 and VIPeR datasets show that the proposed method can achieve
 favorable accuracy while compared with state-of-the-arts.
****

A Multi-Level Contextual Model For Person Recognition in Photo Albums
Haoxiang Li, Jonathan Brandt, Zhe Lin, Xiaohui Shen, Gang Hua; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1
297-1305
In this work, we present a new framework for person recognition in photo albums
 that exploits contextual cues at multiple levels, spanning individual persons, i
ndividual photos, and photo groups. Through experiments, we show that the infor
mation available at each of these distinct contextual levels provides complement
ary cues as to person identities. At the person level, we leverage clothing and
 body appearance in addition to facial appearance, and to compensate for instanc
es where the faces are not visible. At the photo level we leverage a learned pr
ior on the joint distribution of identities on the same photo to guide the ident
ity assignments. Going beyond a single photo, we are able to infer natural group
ings of photos with shared context in an unsupervised manner. By exploiting this
 shared contextual information, we are able to reduce the identity search space
and exploit higher intra-personal appearance consistency within photo groups. Ou
r new framework enables efficient use of these complementary multi-level context
ual cues to improve overall recognition rates on the photo album person recognit
ion task, as demonstrated through state-of-the-art results on a challenging publ
ic dataset. Our results outperform competing methods by a significant margin, w
hile being computationally efficient and practical in a real world application.
****

Unsupervised Cross-Dataset Transfer Learning for Person Re-Identification
Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun H
uang, Yonghong Tian; Proceedings of the IEEE Conference on Computer Vision and P
attern Recognition (CVPR), 2016, pp. 1306-1315
Most existing person re-identification (Re-ID) approaches follow a supervised le
arning framework, in which a large number of labelled matching pairs are require
d for training. This severely limits their scalability in real-world application
s. To overcome this limitation, we develop a novel cross-dataset transfer learni
ng approach to learn a discriminative representation. It is unsupervised in the
sense that the target dataset is completely unlabelled. Specifically, we present
 an multi-task dictionary learning method which is able to learn a dataset-share
d but target-data-biased representation. Experimental results on five benchmark
datasets demonstrate that the method significantly outperforms the state-of-the-
art.
****

Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry
Jiale Cao, Yanwei Pang, Xuelong Li; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2016, pp. 1316-1324
The discrimination and simplicity of features are very important for effective a
nd efficient pedestrian detection. However, most state-of-the-art methods are un
able to achieve good tradeoff between accuracy and efficiency. Inspired by some
simple inherent attributes of pedestrians (i.e., appearance constancy and shape
symmetry), we propose two new types of non-neighboring features (NNF): side-inne
r difference features (SIDF) and symmetrical similarity features (SSF). SIDF can
 characterize the difference between the background and pedestrian and the diffe
rence between the pedestrian contour and its inner part. SSF can capture the sym
metrical similarity of pedestrian shape. However, it's difficult for neighboring
 features to have such above characterization abilities. Finally, we propose to
combine both non-neighboring and neighboring features for pedestrian detection.
It's found that nonneighboring features can further decrease the average miss ra
te by 4.44%. Experimental results on INRIA and Caltech pedestrian datasets demon
strate the effectiveness and efficiency of the proposed method. Compared to the
state-ofthe- art methods without using CNN, our method achieves the best detecti

on performance on Caltech, outperforming the second best method (i.e., Checkboards) by 1.63%.

****

Recurrent Convolutional Network for Video-Based Person Re-Identification
Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1325-1334

In this paper we propose a novel recurrent neural network architecture for video-based person re-identification. Given the video sequence of a person, features are extracted from each frame using a convolutional neural network that incorporates a recurrent final layer, which allows information to flow between time-steps. The features from all time-steps are then combined using temporal pooling to give an overall appearance feature for the complete sequence. The convolutional network, recurrent layer, and temporal pooling layer, are jointly trained to act as a feature extractor for video-based re-identification using a Siamese network architecture. Our approach makes use of colour and optical flow information in order to capture appearance and motion information which is useful for video re-identification. Experiments are conduced on the iLIDS-VID and PRID-2011 datasets to show that this approach outperforms existing methods of video-based re-identification.

****

Person Re-Identification by Multi-Channel Parts-Based CNN With Improved Triplet Loss Function
De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335-1344

Person re-identification across cameras remains a very challenging problem, especially when there are no overlapping fields of view between cameras. In this paper, we present a novel multi-channel parts-based convolutional neural network (CNN) model under the triplet framework for person re-identification. Specifically, the proposed CNN model consists of multiple channels to jointly learn both the global full body and local body-parts features of the input persons. The CNN model is trained by an improved triplet loss function that serves to pull the instances of the same person closer, and at the same time push the instances belonging to different persons farther from each other in the learned feature space. Extensive comparative evaluations demonstrate that our proposed method significantly outperforms many state-of-the-art approaches, including both traditional and deep network-based ones, on the challenging i-LIDS, VIPeR, PRID2011 and CUHK01 datasets.

****

Top-Push Video-Based Person Re-Identification
Jinjie You, Ancong Wu, Xiang Li, Wei-Shi Zheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1345-1353

Most existing person re-identification (re-id) models focus on matching still person images across disjoint camera views using the setting of either single-shot or multi-shot. Since limited information can be exploited from still images, it is hard (if not impossible) to overcome the occlusion, pose and camera-view change, and lighting variation problems. In comparison, video-based re-id methods can utilize extra space-time information, which contains much more rich cues for matching to overcome the mentioned problems. However, in this work, we find that when using video-based representation, some inter-class difference can be much more obscure than the one when using still-image-based representation, because different people could not only have similar appearance but also may have similar motions and actions which are hard to align. To solve this problem, we propose a top-push distance learning model (TDL), in which we integrate a top-push constrain, for matching video features of persons. The top-push constraint enforces the optimization on top-rank matching in re-id, so as to make the matching model more effective towards selecting more discriminative features to distinguish different persons. Our experiments show that the proposed video-based re-id framework outperforms the state-of-the-art video-based re-id methods.

****
Improving Person Re-Identification via Pose-Aware Multi-Shot Matching
Yeong-Jun Cho, Kuk-Jin Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1354-1362
Person re-identification is the problem of recognizing people across images or videos from non-overlapping views. Although there has been much progress in person re-identification for the last decade, it still remains a challenging task because of severe appearance changes of a person due to diverse camera viewpoints and person poses. In this paper, we propose a novel framework for person re-identification by analyzing camera viewpoints and person poses, so-called Pose-aware Multi-shot Matching (PaMM), which robustly estimates target poses and efficiently conducts multi-shot matching based on the target pose information. Experimental results using public person re-identification dataset show that the proposed methods are promising for person re-identification under diverse viewpoints and pose variances.
****
Hierarchical Gaussian Descriptor for Person Re-Identification
Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1363-1372
Describing the color and textural information of a person image is one of the most crucial aspects of person re-identification. In this paper, we present a novel descriptor based on a hierarchical distribution of pixel features. A hierarchical covariance descriptor has been successfully applied for image classification. However, the mean information of pixel features, which is absent in covariance, tends to be major discriminative information of person images. To solve this problem, we describe a local region in an image via hierarchical Gaussian distribution in which both means and covariances are included in their parameters. More specifically, we model the region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch. The characteristics of the set of Gaussians are again described by another Gaussian distribution. In both steps, unlike the hierarchical covariance descriptor, the proposed descriptor can model both the mean and the covariance information of pixel features properly. The results of experiments conducted on five databases indicate that the proposed descriptor exhibits remarkably high performance which outperforms the state-of-the-art descriptors for person re-identification.
****
STCT: Sequentially Training Convolutional Networks for Visual Tracking
Lijun Wang, Wanli Ouyang, Xiaogang Wang, Huchuan Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1373-1381
Due to the limited amount of training samples, fine-tuning pre-trained deep models online is prone to over-fitting. In this paper, we propose a sequential training method for convolutional neural networks (CNNs) to effectively transfer pre-trained deep features for online applications. We regard a CNN as an ensemble with each channel of the output feature map as an individual base learner. Each base learner is trained using different loss criterions to reduce correlation and avoid over-training. To achieve the best ensemble online, all the base learners are sequentially sampled into the ensemble via important sampling. To further improve the robustness of each base learner, we propose to train the convolutional layers with random binary masks, which serves as a regularization to enforce each base learner to focus on different input features. The proposed online training method is applied to visual tracking problem by transferring deep features trained on massive annotated visual data and is shown to significantly improve tracking performance. Extensive experiments are conducted on two challenging benchmark data set and demonstrate that our tracking algorithm can outperform state-of-the-art methods with a considerable margin.
****
Determining Occlusions From Space and Time Image Reconstructions
Juan-Manuel Perez-Rua, Tomas Crivelli, Patrick Bouthemy, Patrick Perez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20

16, pp. 1382-1391
The problem of localizing occlusions between consecutive frames of a video is im portant but rarely tackled on its own. In most works, it is tightly interleaved with the computation of accurate optical flows, which leads to a delicate chicke n-and-egg problem. With this in mind, we propose a novel approach to occlusion d etection where visibility or not of a point in next frame is formulated in terms of visual reconstruction. The key issue is now to determine how well a pixel in the first image can be "recon- structed" from co-located colors in the next ima ge. We first exploit this reasoning at the pixel level with a new detection crit erion. Contrary to the ubiquitous displaced-frame-difference and forward-backwar d flow vector matching, the proposed alternative does not critically depend on a precomputed, dense displacement field, while being shown to be more effective. We then leverage this local modeling within an energy-minimization framework tha t delivers occlusion maps. An easy-to-obtain collection of parametric motion mod els is exploited within the energy to provide the required level of motion infor mation. Our approach outperforms state-of-the-art detection methods on the chall enging MPI Sintel dataset.
****

Online Multi-Object Tracking via Structural Constraint Event Aggregation
Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, Kuk-Jin Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 13 92-1400
Multi-object tracking (MOT) becomes more challenging when objects of interest ha ve similar appearances. In that case, the motion cues are particularly useful fo r discriminating multiple objects. However, for online 2D MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera mo vements and thus not always smooth or predictable. To deal with such unexpected camera motion for online 2D MOT, a structural motion constraint between objects has been utilized thanks to its robustness to camera motion. In this paper, we p ropose a new data association method that effectively exploits structural motion constraints in the presence of large camera motion. In addition, to further imp rove the robustness of data association against mis-detections and clutters, a n ovel event aggregation approach is developed to integrate structural constraints in assignment costs for online MOT. Experimental results on a large number of d atasets demonstrate the effectiveness of the proposed algorithm for online 2D MO T.
****

Staple: Complementary Learners for Real-Time Tracking
Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, Philip H. S. Tor r; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1401-1409
Correlation Filter-based trackers have recently achieved excellent performance, showing great robustness to challenging situations exhibiting motion blur and il lumination changes. However, since the model that they learn depends strongly on the spatial layout of the tracked object, they are notoriously sensitive to def ormation. Models based on colour statistics have complementary traits: they cope well with variation in shape, but suffer when illumination is not consistent th roughout a sequence. Moreover, colour distributions alone can be insufficiently discriminative. In this paper, we show that a simple tracker combining complemen tary cues in a ridge regression framework can operate faster than 80 FPS and out perform not only all entries in the popular VOT14 competition, but also recent a nd far more sophisticated trackers according to multiple benchmarks.
****

Robust Optical Flow Estimation of Double-Layer Images Under Transparency or Refl ection
Jiaolong Yang, Hongdong Li, Yuchao Dai, Robby T. Tan; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1410-1419
This paper deals with a challenging, frequently encountered, yet not properly in vestigated problem in two-frame optical flow estimation. That is, the input fram es are compounds of two imaging layers -- one desired background layer of the sc

ene, and one distracting, possibly moving layer due to transparency or reflectio n. In this situation, the conventional brightness constancy constraint -- the co rnerstone of most existing optical flow methods -- will no longer be valid. In t his paper, we propose a robust solution to this problem. The proposed method per forms both optical flow estimation, and image layer separation. It exploits a ge neralized double-layer brightness consistency constraint connecting these two ta sks, and utilizes the priors for both of them. Experiments on both synthetic dat a and real images have confirmed the efficacy of the proposed method. To the bes t of our knowledge, this is the first attempt towards handling generic optical f low fields of two-frame images containing transparency or reflection.
****

Siamese Instance Search for Tracking
Ran Tao, Efstratios Gavves, Arnold W.M. Smeulders; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1420-1429
In this paper we present a tracker, which is radically different from state-of-t he-art trackers: we apply no model updating, no occlusion detection, no combinat ion of trackers, no geometric matching, and still deliver state-of-the-art track ing performance, as demonstrated on the popular online tracking benchmark (OTB) and six very challenging YouTube videos. The presented tracker simply matches th e initial patch of the target in the first frame with candidates in a new frame and returns the most similar patch by a learned matching function. The strength of the matching function comes from being extensively trained generically, i.e., without any data of the target, using a Siamese deep neural network, which we d esign for tracking. Once learned, the matching function is used as is, without a ny adapting, to track previously unseen targets. It turns out that the learned m atching function is so powerful that a simple tracker built upon it, coined Siam ese INstance search Tracker, SINT, which only uses the original observation of t he target from the first frame, suffices to reach state-of-the-art performance. Further, we show the proposed tracker even allows for target re-identification a fter the target was absent for a complete video shot.
****

Adaptive Decontamination of the Training Set: A Unified Formulation for Discrimi native Visual Tracking
Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, Michael Felsberg; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 , pp. 1430-1438
Tracking-by-detection methods have demonstrated competitive performance in recen t years. In these approaches, the tracking model heavily relies on the quality o f the training set. Due to the limited amount of labeled training data, addition al samples need to be extracted and labeled by the tracker itself. This often le ads to the inclusion of corrupted training samples, due to occlusions, misalignm ents and other perturbations. Existing tracking-by-detection methods either igno re this problem, or employ a separate component for managing the training set. We propose a novel generic approach for alleviating the problem of corrupted tr aining samples in tracking-by-detection frameworks. Our approach dynamically man ages the training set by estimating the quality of the samples. Contrary to exis ting approaches, we propose a unified formulation by minimizing a single loss ov er both the target appearance model and the sample quality weights. The joint fo rmulation enables corrupted samples to be down-weighted while increasing the imp act of correct ones. Experiments are performed on three benchmarks: OTB-2015 wit h 100 videos, VOT-2015 with 60 videos, and Temple-Color with 128 videos. On the OTB-2015, our unified formulation significantly improves the baseline, with a ga in of 3.8% in mean overlap precision. Finally, our method achieves state-of-the- art results on all three datasets.
****

3D Part-Based Sparse Tracker With Automatic Synchronization and Registration
Adel Bibi, Tianzhu Zhang, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1439-1448
In this paper, we present a part-based sparse tracker in a particle filter frame work where both the motion and appearance model are formulated in 3D. The motion

model is adaptive and directed according to a simple yet powerful occlusion handling paradigm, which is intrinsically fused in the motion model. Also, since 3D trackers are sensitive to synchronization and registration noise in the RGB and depth streams, we propose automated methods to solve these two issues. Extensive experiments are conducted on a popular RGBD tracking benchmark, which demonstrate that our tracker can achieve superior results, outperforming many other recent and state-of-the-art RGBD trackers.

****

Recurrently Target-Attending Tracking

Zhen Cui, Shengtao Xiao, Jiashi Feng, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1449-1458

Robust visual tracking is a challenging task in computer vision. Due to the accumulation and propagation of estimation error, model drifting often occurs and degrades the tracking performance. To mitigate this problem, in this paper we propose a novel tracking method called Recurrently Target-attending Tracking (RTT). RTT attempts to identify and exploit those reliable parts which are beneficial for the overall tracking process. To bypass occlusion and discover reliable components, multi-directional Recurrent Neural Networks (RNNs) are employed in RTT to capture long-range contextual cues by traversing a candidate spatial region from multiple directions. The produced confidence maps from the RNNs are employed to adaptively regularize the learning of discriminative correlation filters by suppressing clutter background noises while making full use of the information from reliable parts. To solve the weighted correlation filters, we especially derive an efficient closed-form solution with a sharp reduction in computation complexity. Extensive experiments demonstrate that our proposed RTT is more competitive over those correlation filter based methods.

****

Structured Regression Gradient Boosting

Ferran Diego, Fred A. Hamprecht; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1459-1467

We propose a new way to train a structured output prediction model. More specifically, we train nonlinear data terms in a Gaussian Conditional Random Field (GCRF) by a generalized version of gradient boosting. The approach is evaluated on three challenging regression benchmarks: vessel detection, single image depth estimation and image inpainting. These experiments suggest that the proposed boosting framework matches or exceeds the state-of-the-art.

****

Loss Functions for Top-k Error: Analysis and Insights

Maksim Lapin, Matthias Hein, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1468-1477

In order to push the performance on realistic computer vision tasks, the number of classes in modern benchmark datasets has significantly increased in recent years. This increase in the number of classes comes along with increased ambiguity between the class labels, raising the question if top-1 error is the right performance measure. In this paper, we provide an extensive comparison and evaluation of established multiclass methods comparing their top-k performance both from a practical as well as from a theoretical perspective. Moreover, we introduce novel top-k loss functions as modifications of the softmax and the multiclass SVM losses and provide efficient optimization schemes for them. In the experiments, we compare on various datasets all of the proposed and established methods for top-k error optimization. An interesting insight of this paper is that the softmax loss yields competitive top-k performance for all k simultaneously. For a specific top-k error, our new top- k losses lead typically to further improvements while being faster to train than the softmax.

****

Metric Learning as Convex Combinations of Local Models With Generalization Guarantees

Valentina Zantedeschi, Remi Emonet, Marc Sebban; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1478-1486

Over the past ten years, metric learning allowed the improvement of the numerous

machine learning approaches that manipulate distances or similarities. In this field, local metric learning has been shown to be very efficient, especially to take into account non linearities in the data and better capture the peculiarities of the application of interest. However, it is well known that local metric learning (i) can entail overfitting and (ii) face difficulties to compare two instances that are assigned to two different local models. In this paper, we address these two issues by introducing a novel metric learning algorithm that linearly combines local models (C2LM). Starting from a partition of the space in regions and a model (a score function) for each region, C2LM defines a metric between points as a weighted combination of the models. A weight vector is learned for each pair of regions, and a spatial regularization ensures that the weight vectors evolve smoothly and that nearby models are favored in the combination. The proposed approach has the particularity of working in a regression setting, of working implicitly at different scales, and of being generic enough so that it is applicable to similarities and distances. We prove theoretical guarantees of the approach using the framework of algorithmic robustness. We carry out experiments with datasets using both distances (perceptual color distances, using Mahalanobis-like distances) and similarities (semantic word similarities, using bilinear forms), showing that C2LM consistently improves regression accuracy even in the case where the amount of training data is small.

****

Efficient Training of Very Deep Neural Networks for Supervised Hashing

Ziming Zhang, Yuting Chen, Venkatesh Saligrama; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1487-1495

In this paper, we propose training very deep neural networks (DNNs) for supervised learning of hash codes. Existing methods in this context train relatively "shallow" networks limited by the issues arising in back propagation (e.g. vanishing gradients) as well as computational efficiency. We propose a novel and efficient training algorithm inspired by alternating direction method of multipliers (ADMM) that overcomes some of these limitations. Our method decomposes the training process into independent layer-wise local updates through auxiliary variables. Empirically we observe that our training algorithm always converges and its computational complexity is linearly proportional to the number of edges in the networks. Empirically we manage to train DNNs with 64 hidden layers and 1024 nodes per layer for supervised hashing in about 3 hours using a single GPU. Our proposed very deep supervised hashing (VDSH) method significantly outperforms the state-of-the-art on several benchmark datasets.

****

Information Bottleneck Learning Using Privileged Information for Visual Recognition

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, Gianfranco Doretto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1496-1505

We explore the visual recognition problem from a main data view when an auxiliary data view is available during training. This is important because it allows improving the training of visual classifiers when paired additional data is cheaply available, and it improves the recognition from multi-view data when there is a missing view at testing time. The problem is challenging because of the intrinsic asymmetry caused by the missing auxiliary view during testing. We account for such view during training by extending the information bottleneck method, and by combining it with risk minimization. In this way, we establish an information theoretic principle for leaning any type of visual classifier under this particular setting. We use this principle to design a large-margin classifier with an efficient optimization in the primal space. We extensively compare our method with the state-of-the-art on different visual recognition datasets, and with different types of auxiliary data, and show that the proposed framework has a very promising potential.

****

3D Action Recognition From Novel Viewpoints

Hossein Rahmani, Ajmal Mian; Proceedings of the IEEE Conference on Computer Visi

on and Pattern Recognition (CVPR), 2016, pp. 1506-1515
We propose a human pose representation model that transfers human poses acquired from different unknown views to a view-invariant high-level space. The model is a deep convolutional neural network and requires a large corpus of multiview training data which is very expensive to acquire. Therefore, we propose a method to generate this data by fitting synthetic 3D human models to real motion capture data and rendering the human poses from numerous viewpoints. While learning the CNN model, we do not use action labels but only the pose labels after clustering all training poses into k clusters. The proposed model is able to generalize to real depth images of unseen poses without the need for re-training or fine-tuning. Real depth videos are passed through the model frame-wise to extract view-invariant features. For spatio-temporal representation, we propose group sparse Fourier Temporal Pyramid which robustly encodes the action specific most discriminative output features of the proposed human pose model. Experiments on two multiview and three single-view benchmark datasets show that the proposed method dramatically outperforms existing state-of-the-art in action recognition.
****

3D Shape Attributes
David F. Fouhey, Abhinav Gupta, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1516-1524
In this paper we investigate 3D attributes as a means to understand the shape of an object in a single image. To this end, we make a number of contributions: (i) we introduce and define a set of 3D Shape attributes, including planarity, symmetry and occupied space; (ii) we show that such properties can be successfully inferred from a single image using a Convolutional Neural Network (CNN); (iii) we introduce a 143K image dataset of sculptures with 2197 works over 242 artists for training and evaluating the CNN; (iv) we show that the 3D attributes trained on this dataset generalize to images of other (non-sculpture) object classes; and furthermore (v) we show that the CNN also provides a shape embedding that can be used to match previously unseen sculptures largely independent of viewpoint.
****

Three-Dimensional Object Detection and Layout Prediction Using Clouds of Oriented Gradients
Zhile Ren, Erik B. Sudderth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1525-1533
We develop new representations and algorithms for three-dimensional (3D) object detection and spatial layout prediction in cluttered indoor scenes. RGB-D images are traditionally described by local geometric features of the 3D point cloud. We propose a cloud of oriented gradient (COG) descriptor that links the 2D appearance and 3D pose of object categories, and thus accurately models how perspective projection affects perceived image boundaries. We also propose a "Manhattan voxel" representation which better captures the 3D room layout geometry of common indoor environments. Effective classification rules are learned via a structured prediction framework that accounts for the intersection-over-union overlap of hypothesized 3D cuboids with human annotations, as well as orientation estimation errors. Contextual relationships among categories and layout are captured via a cascade of classifiers, leading to holistic scene hypotheses with improved accuracy. Our model is learned solely from annotated RGB-D images, without the benefit of CAD models, but nevertheless its performance substantially exceeds the state-of-the-art on the SUN RGB-D database. Avoiding CAD models allows easier learning of detectors for many object categories.
****

3D Semantic Parsing of Large-Scale Indoor Spaces
Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1534-1543
In this paper, we propose a method for semantic parsing the 3D point cloud of an entire building using a hierarchical approach: first, the raw data is parsed into semantically meaningful spaces (e.g. rooms, etc) that are aligned into a cano

nical reference coordinate system. Second, the spaces are parsed into their structural and building elements (e.g. walls, columns, etc). Performing these with a strong notation of global 3D space is the backbone of our method. The alignment in the first step injects strong 3D priors from the canonical coordinate system into the second step for dis- covering elements. This allows diverse challenging scenarios as man-made indoor spaces often show recurrent geo- metric patterns while the appearance features can change drastically. We also argue that identification of structural elements in indoor spaces is essentially a detection problem, rather than segmentation which is commonly used. We evaluated our method on a new dataset of several buildings with a covered area of over 6, 000m2 and over 215 million points, demonstrating robust results readily useful for practical applications.

****

Dense Human Body Correspondences Using Convolutional Networks

Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, Hao Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1544-1553

We propose a deep learning approach for finding dense correspondences between 3D scans of people. Our method requires only partial geometric information in the form of two depth maps or partial reconstructed surfaces, works for humans in arbitrary poses and wearing any clothing, does not require the two people to be scanned from similar viewpoints, and runs in real time. We use a deep convolutional neural network to train a feature descriptor on depth map pixels, but crucially, rather than training the network to solve the shape correspondence problem directly, we train it to solve a body region classification problem, modified to increase the smoothness of the learned descriptors near region boundaries. This approach ensures that nearby points on the human body are nearby in feature space, and vice versa, rendering the feature descriptor suitable for computing dense correspondences between the scans. We validate our method on real and synthetic data for both clothed and unclothed humans, and show that our correspondences are more robust than is possible with state-of-the-art unsupervised methods, and more accurate to those found using methods that require full watertight 3D geometry.

****

Geometry-Informed Material Recognition

Joseph DeGol, Mani Golparvar-Fard, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1554-1562

Our goal is to recognize material categories using images and geometry information. In many applications, such as construction management, coarse geometry information is available. We investigate how 3D geometry (surface normals, camera intrinsic and extrinsic parameters) can be used with 2D features (texture and color) to improve material classification. We introduce a new dataset, GeoMat, which is the first to provide both image and geometry data in the form of: (i) training and testing patches that were extracted at different scales and perspectives from real world examples of each material category, and (ii) a large scale construction site scene that includes 160 images and over 800,000 hand labeled 3D points. Our results show that using 2D and 3D features both jointly and independently to model materials improves classification accuracy across multiple scales and viewing directions for both material patches and images of a large scale construction site scene.

****

Towards Open Set Deep Networks

Abhijit Bendale, Terrance E. Boult; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1563-1572

Deep networks have produced significant gains for various visual recognition problems, leading to high impact academic and commercial applications. Recent work in deep networks highlighted that it is easy to generate images that humans would never classify as a particular object class, yet networks classify such images high confidence as that given class - deep network are easily fooled with images humans do not consider meaningful. The closed set nature of deep networks fo

rces them to choose from one of the known classes leading to such artifacts. Recognition in the real world is open set, i.e. the recognition system should reject unknown/unseen classes at test time. We present a methodology to adapt deep networks for open set recognition, by introducing a new model layer, OpenMax, which estimates the probability of an input being from an unknown class. A key element of estimating the unknown probability is adapting Meta-Recognition concepts to the activation patterns in the penultimate layer of the network. OpenMax allows rejection of "fooling" and unrelated open set images presented to the system; OpenMax greatly reduces the number of obvious errors made by a deep network. We prove that the OpenMax concept provides bounded open space risk, thereby formally providing an open set recognition solution. We evaluate the resulting open set deep networks using pre-trained networks from the Caffe Model-zoo on ImageNet 2012 validation data, and thousands of fooling and open set images. The proposed OpenMax model significantly outperforms open set recognition accuracy of basic deep networks as well as deep networks with thresholding of SoftMax probabilities.

****

What's Wrong With That Object? Identifying Images of Unusual Objects by Modelling the Detection Score Distribution

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 Recognition (CVPR), 2016, pp. 1573-1581

This paper studies the challenging problem of identifying unusual instances of known objects in images within an "open world" setting. That is, we aim to find objects that are members of a known class, but which are not typical of that class. Thus the "unusual object" should be distinguished from both the "regular object" and the "other objects". Such unusual objects may be of interest in many applications such as surveillance or quality control. We propose to identify unusual objects by inspecting the distribution of object detection scores at multiple image regions. The key observation motivating our approach is that "regular object" images, "unusual object" images and "other objects" images exhibit different region-level scores in terms of both the score values and the spatial distributions. To model these distributions we propose to use Gaussian Processes (GP) to construct two separate generative models, one for the "regular object" and the other for the "other objects". More specifically, we design a new covariance function to simultaneously model the detection score at a single location and the score dependencies between multiple regions. We demonstrate that the proposed approach outperforms comparable methods on a new large dataset constructed for the purpose.

****

Large-Scale Location Recognition and the Geometric Burstiness Problem

Torsten Sattler, Michal Havlena, Konrad Schindler, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1582-1590

Visual location recognition is the task of determining the place depicted in a query image from a given database of geo-tagged images. Location recognition is often cast as an image retrieval problem and recent research has almost exclusively focused on improving the chance that a relevant database image is ranked high enough after retrieval. The implicit assumption is that the number of inliers found by spatial verification can be used to distinguish between a related and an unrelated database photo with high precision. In this paper, we show that this assumption does not hold for large datasets due to the appearance of geometric bursts, i.e., sets of visual elements appearing in similar geometric configurations in unrelated database photos. We propose algorithms for detecting and handling geometric bursts. Although conceptually simple, using the proposed weighting schemes dramatically improves the recall that can be achieved when high precision is required compared to the standard re-ranking based on the inlier count. Our approach is easy to implement and can easily be integrated into existing location recognition systems

****

Regularity-Driven Facade Matching Between Aerial and Street Views
Mark Wolff, Robert T. Collins, Yanxi Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1591-1600
We present an approach for detecting and matching building facades between aerial view and street-view images. We exploit the regularity of urban scene facades as captured by their lattice structures and deduced from median-tiles' shape context, color, texture and spatial similarities. Our experimental results demonstrate effective matching of oblique and partially-occluded facades between aerial and ground views. Quantitative comparisons for automated urban scene facade matching from three cities show superior performance of our method over baseline SIFT, Root-SIFT and the more sophisticated Scale-Selective Self-Similarity and Binary Coherent Edge descriptors. We also illustrate regularity-based applications of occlusion removal from street views and higher-resolution texture-replacement in aerial views.
****

Do Computational Models Differ Systematically From Human Object Perception?
R. T. Pramod, S. P. Arun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1601-1609
Recent advances in neural networks have revolutionized computer vision, but these algorithms are still outperformed by humans. Could this performance gap be due to systematic differences between object representations in humans and machines? To answer this question we collected a large dataset of 26,675 perceived dissimilarity measurements from 2,801 visual objects across 269 human subjects, and used this dataset to train and test leading computational models. The best model (a combination of all models) accounted for 68% of the explainable variance. Importantly, all computational models showed systematic deviations from perception: (1) They underestimated perceptual distances between objects with symmetry or large area differences; (2) They overestimated perceptual distances between objects with shared features. Our results reveal critical elements missing in computer vision algorithms and point to explicit encoding of these properties in higher visual areas in the brain.
****

Contour Detection in Unstructured 3D Point Clouds
Timo Hackel, Jan D. Wegner, Konrad Schindler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1610-1618
We describe a method to automatically detect contours, i.e. lines along which the surface orientation sharply changes, in large-scale outdoor point clouds. Contours are important intermediate features for structuring point clouds and converting them into high-quality surface or solid models, and are extensively used in graphics and mapping applications. Yet, detecting them in unstructured, inhomogeneous point clouds turns out to be surprisingly difficult, and existing line detection algorithms largely fail. We approach contour extraction as a two-stage discriminative learning problem. In the first stage, a contour score for each individual point is predicted with a binary classifier, using a set of features extracted from the point's neighborhood. The contour scores serve as a basis to construct an overcomplete graph of candidate contours. The second stage selects an optimal set of contours from the candidates. This amounts to a further binary classification in a higher-order MRF, whose cliques encode a preference for connected contours and penalize loose ends. The method can handle point clouds >10^7 points in a couple of minutes, and vastly outperforms a baseline that performs Canny-style edge detection on a range image representation of the point cloud.
****

Unsupervised Learning of Edges
Yin Li, Manohar Paluri, James M. Rehg, Piotr Dollar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1619-1627
Data-driven approaches for edge detection have proven effective and achieve top results on modern benchmarks. However, all current data-driven edge detectors require manual supervision for training in the form of hand-labeled region segments or object boundaries. Specifically, human annotators mark semantically meaningful edges which are subsequently used for training. Is this form of strong, high

-level supervision actually necessary to learn to accurately detect edges? In th is work we present a simple yet effective approach for training edge detectors w ithout human supervision. To this end we utilize motion, and more specifically, the only input to our method is noisy semi-dense matches between frames. We begi n with only a rudimentary knowledge of edges (in the form of image gradients), a nd alternate between improving motion estimation and edge detection in turn. Usi ng a large corpus of video data, we show that edge detectors trained using our u nsupervised scheme approach the performance of the same methods trained with ful l supervision (within 3-5%). Finally, we show that when using a deep network for the edge detector, our approach provides a novel pre-training scheme for object detection.

****

Blind Image Deblurring Using Dark Channel Prior

Jinshan Pan, Deqing Sun, Hanspeter Pfister, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 162 8-1636

We present a simple and effective blind image deblurring method based on the dar k channel prior. Our work is inspired by the interesting observation that the da rk channel of blurred images is less sparse. While most image patches in the cle an image contain some dark pixels, these pixels are not dark when averaged with neighboring high-intensity pixels during the blur process.Our analysis shows tha t this change in the sparsity of the dark channel is an inherent property of the blur process, both theoretically and empirically. This change in the sparsity o f the dark channel is an inherent property of the blur process, which we both pr ove mathematically and validate using training data. Therefore, enforcing the sp arsity of the dark channel helps blind deblurring on various scenarios, includin g natural, face, text, and low-illumination images. However, sparsity of the dar k channel introduces a non-convex non-linear optimization problem. We introduce a linear approximation of the min operator to compute the dark channel. Our look -up-table-based method converges fast in practice and can be directly extended t o non-uniform deblurring. Extensive experiments show that our method achieves st ate-of-the-art results on deblurring natural images and compares favorably metho ds that are well-engineered for specific scenarios.

****

Deeply-Recursive Convolutional Network for Image Super-Resolution

Jiwon Kim, Jung Kwon Lee, Kyoung Mu Lee; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1637-1645

We propose an image super-resolution method (SR) using a deeply-recursive convol utional network (DRCN). Our network has a very deep recursive layer (up to 16 re cursions). Increasing recursion depth can improve performance without introducin g new parameters for additional convolutions. Albeit advantages, learning a DRCN is very hard with a standard gradient descent method due to exploding/ vanishin g gradients. To ease the difficulty of training, we propose two extensions: recu rsive supervision and skip-connection. Our method outperforms previous methods b y a large margin.

****

Accurate Image Super-Resolution Using Very Deep Convolutional Networks

Jiwon Kim, Jung Kwon Lee, Kyoung Mu Lee; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1646-1654

We present a highly accurate single image superresolution (SR) method. Our metho d uses a very deep convolutional network inspired by VGG-net used for ImageNet c lassification [19]. We find increasing our network depth shows a significant imp rovement in accuracy. Our final model uses 20 weight layers. By cascading small filters many times in a deep network structure, contextual information over larg e image regions is exploited in an efficient way. With very deep networks, howev er, convergence speed becomes a critical issue during training. We propose a sim ple yet effective training procedure. We learn residuals only and use extremely high learning rates (104 times higher than SRCNN [6]) enabled by adjustable grad ient clipping. Our proposed method performs better than existing methods in accu racy and visual improvements in our results are easily noticeable.

****
RAW Image Reconstruction Using a Self-Contained sRGB-JPEG Image With Only 64 KB Overhead
Rang M. H. Nguyen, Michael S. Brown; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2016, pp. 1655-1663
Most camera images are saved as 8-bit standard RGB (sRGB) compressed JPEGs. Eve n when JPEG compression is set to its highest quality, the encoded sRGB image ha s been significantly processed in terms of color and tone manipulation. This m akes sRGB-JPEG images undesirable for many computer vision tasks that assume a d irect relationship between pixel values and incoming light. For such applicatio ns, the RAW image format is preferred, as RAW represents a minimally processed, sensor-specific RGB image with higher dynamic range that is linear with respect to scene radiance. The drawback with RAW images, however, is that they require large amounts of storage and are not well-supported by many imaging application s. To address this issue, we present a method to encode the necessary metadata within an sRGB image to reconstruct a high-quality RAW image. Our approach re quires no calibration of the camera and can reconstruct the original RAW to with in 0.3% error with only a 64 KB overhead for the additional data. More importa ntly, our output is a fully self-contained 100% complainant sRGB-JPEG file that can be used as-is, not affecting any existing image workflow - the RAW image can be extracted when needed, or ignored otherwise. We detail our approach and sho w its effectiveness against competing strategies.
****
Group MAD Competition - A New Methodology to Compare Objective Image Quality Mod els
Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogn ition (CVPR), 2016, pp. 1664-1673
Objective image quality assessment (IQA) models aim to automatically predict hum an visual perception of image quality and are of fundamental importance in the f ield of image processing and computer vision. With an increasing number of IQA m odels proposed, how to fairly compare their performance becomes a major challeng e due to the enormous size of the image space and the limited resource for subje ctive testing. The standard approach in the literature is to compute several cor relation metrics between subjective mean opinion scores (MOSs) and objective mod el predictions on several well-known subject-rated databases that contain distor ted images generated from a few dozens of source images, which provide an extrem ely limited representation of real-world images. Moreover, most IQA models were developed after these databases became publicly available and often involve mach ine learning or manual parameter tuning steps to boost their performance on thes e databases, and thus their generalization capabilities are questionable. Here w e propose a substantially different methodology to compare IQA models. We first build a database that contains 4,744 source natural images, together with 94,880 distorted images created from them. We then propose a novel mechanism, namely g roup MAximum Differentiation (gMAD) competition, that helps automatically select subsets of image pairs from the database that provide the strongest test to let the IQA models compete with each other. Subjective testing on the selected subs ets reveals the relative performance of the IQA models and provides useful insig hts on potential ways to improve them. We report the gMAD competition results be tween 16 well-known IQA models, but the framework is extendable, allowing future IQA models to be added into the competition.
****
Non-Local Image Dehazing
Dana Berman, Tali treibitz, Shai Avidan; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1674-1682
Haze limits visibility and reduces image contrast in outdoor images. The degrada tion is different for every pixel and depends on the distance of the scene point from the camera. This dependency is expressed in the transmission coefficients, that control the scene attenuation and amount of haze in every pixel. Previous methods solve the single image dehazing problem using various patch-based priors

. We, on the other hand, propose an algorithm based on a new, non-local prior. T
he algorithm relies on the assumption that colors of a haze-free image are well
approximated by a few hundred distinct colors, that form tight clusters in RGB s
pace. Our key observation is that pixels in a given cluster are often non-local,
 i.e., they are spread over the entire image plane and are located at different
distances from the camera. In the presence of haze these varying distances trans
late to different transmission coefficients. Therefore, each color cluster in th
e clear image becomes a line in RGB space, that we term a haze-line. Using these
 haze-lines, our algorithm recovers both the distance map and the haze-free imag
e. The algorithm is linear in the size of the image, deterministic and requires
no training. It performs well on a wide variety of images and is competitive wit
h other state-of-the-art methods.
****

A Holistic Approach to Cross-Channel Image Noise Modeling and Its Application to
 Image Denoising
Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, Seon Joo Kim; Proceedings o
f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p
p. 1683-1691
Modelling and analyzing noise in images is a fundamental task in many computer v
ision systems. Traditionally, noise has been modelled per color channel assuming
 that the color channels are independent. Although the color channels can be con
sidered as mutually independent in camera RAW images, signals from different col
or channels get mixed during the imaging process inside the camera due to gamut
mapping, tone-mapping, and compression. We show the influence of the in-camera i
maging pipeline on noise and propose a new noise model in the 3D RGB space to ac
counts for the color channel mix-ups. A data-driven approach for determining the
 parameters of the new noise model is introduced as well as its application to i
mage denoising. The experiments show that our noise model represents the noise i
n regular JPEG images more accurately compared to the previous models and is adv
antageous in image denoising.
****

Multispectral Images Denoising by Intrinsic Tensor Sparsity Regularization
Qi Xie, Qian Zhao, Deyu Meng, Zongben Xu, Shuhang Gu, Wangmeng Zuo, Lei Zhang; P
roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV
PR), 2016, pp. 1692-1700
Multispectral images (MSI) can help deliver more faithful representation for rea
l scenes than the traditional image system, and enhance the performance of many
computer vision tasks. In real cases, however, an MSI is always corrupted by var
ious noises. In this paper, we propose a new tensor-based denoising approach by
fully considering two intrinsic characteristics underlying an MSI, i.e., the glo
bal correlation along spectrum (GCS) and nonlocal self-similarity across space (
NSS). In specific, we construct a new tensor sparsity measure, called intrinsic
tensor sparsity (ITS) measure, which encodes both sparsity insights delivered by
 the most typical Tucker and CANDECOMP/PARAFAC (CP) low-rank decomposition for a
 general tensor. Then we build a new MSI denoising model by applying the propose
d ITS measure on tensors formed by non-local similar patches within the MSI. The
 intrinsic GCS and NSS knowledge can then be efficiently explored under the regu
larization of this tensor sparsity measure to finely rectify the recovery of a M
SI from its corruption. A series of experiments on simulated and real MSI denois
ing problems show that our method outperforms all state-of-the-arts under compre
hensive quantitative performance measures.
****

A Comparative Study for Single Image Blind Deblurring
Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, Ming-Hsuan Yang; Proceedin
gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201
6, pp. 1701-1709
Numerous single image blind deblurring algorithms have been proposed to restore
latent sharp images under camera motion. However, these algorithms are mainly ev
aluated using either synthetic datasets or few selected real blurred images. It
is thus unclear how these algorithms would perform on images acquired "in the wi

ld" and how we could gauge the progress in the field. In this paper, we aim to b ridge this gap. We present the first comprehensive perceptual study and analysis of single image blind deblurring using real-world blurred images. First, we col lect a dataset of real blurred images and a dataset of synthetically blurred ima ges. Using these datasets, we conduct a large-scale user study to quantify the p erformance of several representative state-of-the-art blind deblurring algorithm s. Second, we systematically analyze subject preferences, including the level of agreement, significance tests of score differences, and rationales for preferri ng one method over another. Third, we study the correlation between human subjec tive scores and several full-reference and no-reference image quality metrics. O ur evaluation and analysis indicate the performance gap between synthetically bl urred images and real blurred image and sheds light on future research in single image blind deblurring.

****

Spatiotemporal Bundle Adjustment for Dynamic 3D Reconstruction

Minh Vo, Srinivasa G. Narasimhan, Yaser Sheikh; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1710-1718

Bundle adjustment jointly optimizes camera intrinsics and extrinsics and 3D poin t triangulation to reconstruct a static scene. The triangulation constraint howe ver is invalid for moving points captured in multiple unsynchronized videos and bundle adjustment is not purposed to estimate the temporal alignment between cam eras. In this paper, we present a spatiotemporal bundle adjustment approach that jointly optimizes four coupled sub-problems: estimating camera intrinsics and e xtrinsics, triangulating 3D static points, as well as subframe temporal alignmen t between cameras and estimating 3D trajectories of dynamic points. Key to our j oint optimization is the careful integration of physics-based motion priors with in the reconstruction pipeline, validated on a large motion capture corpus. We p resent an end-to-end pipeline that takes multiple uncalibrated and unsynchronize d video streams and produces a dynamic reconstruction of the event. Because the videos are aligned with sub-frame precision, we reconstruct 3D trajectories of u nconstrained outdoor activities at much higher temporal resolution than the inpu t videos.

****

Inextensible Non-Rigid Shape-From-Motion by Second-Order Cone Programming

Ajad Chhatkuli, Daniel Pizarro, Toby Collins, Adrien Bartoli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 17 19-1727

We present a global and convex formulation for template-less 3D reconstruction o f a deforming object with the perspective camera. We show for the first time how to construct a Second-Order Cone Programming (SOCP) problem for Non-Rigid Sh ape-from-Motion (NRSfM) using the Maximum-Depth Heuristic (MDH). In this rega rd, we deviate strongly from the general trend of using affine cameras and facto rization-based methods to solve NRSfM. In MDH, the points' depths are maximiz ed so that the distance between neighbouring points in camera space are upper bo unded by the geodesic distance. In NRSfM both geodesic and camera space dista nces are unknown. We show that, nonetheless, given point correspondences and the camera's intrinsics the whole problem is convex and solvable with SOCP. We s how with extensive experiments that our method accurately reconstructs quasi-iso metric surfaces from partial views under articulated and strong deformations. It naturally handles missing correspondences, non-smooth objects and is very simpl e to implement compared to previous methods, with only one free parameter (the n eighbourhood size).

****

Optimal Relative Pose With Unknown Correspondences

Johan Fredriksson, Viktor Larsson, Carl Olsson, Fredrik Kahl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 17 28-1736

Previous work on estimating the epipolar geometry of two views relies on being a ble to reliably match feature points based on appearance. In this paper, we go o ne step further and show that it is feasible to compute both the epipolar geomet

ry and the correspondences at the same time based on geometry only. We do this i
n a globally optimal manner. Our approach is based on an efficient branch and bo
und technique in combination with bipartite matching to solve the correspondence
 problem. We rely on several recent works to obtain good bounding functions to b
attle the combinatorial explosion of possible matchings. It is experimentally de
monstrated that more difficult cases can be handled and that more inlier corresp
ondences can be obtained by being less restrictive in the matching phase.
****

Homography Estimation From the Common Self-Polar Triangle of Separate Ellipses
Haifei Huang, Hui Zhang, Yiu-ming Cheung; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1737-1744
How to avoid ambiguity is a challenging problem for conic-based homography estim
ation. In this paper, we address the problem of homography estimation from two s
eparate ellipses. We find that any two ellipses have a unique common self-polar
triangle, which can provide three line correspondences. Furthermore, by investig
ating the location features of the common self-polar triangle, we show that one
vertex of the triangle lies outside of both ellipses, while the other two vertic
es lies inside the ellipses separately. Accordingly, one more line correspondenc
e can be obtained from the intersections of the conics and the common self-polar
 triangle. Therefore, four line correspondences can be obtained based on the com
mon self-polar triangle, which can provide enough constraints for the homography
 estimation. The main contributions in this paper include: (1) A new discovery o
n the location features of the common self-polar triangle of separate ellipses.
(2) A novel approach for homography estimation. Simulate experiments and real e
xperiments are conducted to demonstrate the feasibility and accuracy of our appr
oach.
****

Heterogeneous Light Fields
Maximilian Diebold, Bernd Jahne, Alexander Gatto; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1745-1753
In contrast to traditional binocular or multi-view stereo approaches, the adequa
tely sampled space of observations in light-field imaging allows, to obtain dens
e and high quality depth maps. It also extends capabilities beyond those of trad
itional methods. Previously, constant intensity has been assumed for estimating
disparity of orientation in most approaches to analyze epipolar plane images (EP
Is). Here, we introduce a modified structure tensor approach which improves dept
h estimation. This extension also includes a model of non-constant intensity on
EPI manifolds. We derive an approach to estimate high quality depth maps in lumi
nance-gradient light fields, as well as in color-filtered light fields. Color-fi
ltered light fields pose particular challenges due to the fact that structures c
an change significantly in appearance with wavelength and can completely vanish
at some wavelength. We demonstrate solutions to this challenge and obtain a dens
e sRGB image reconstruction in addition to dense depth maps.
****

A Consensus-Based Framework for Distributed Bundle Adjustment
Anders Eriksson, John Bastian, Tat-Jun Chin, Mats Isaksson; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1754
-1762
In this paper we study large-scale optimization problems in multi-view geometry,
 in particular the Bundle Adjustment problem. In its conventional formulation, t
he complexity of existing solvers scale poorly with problem size, hence this com
ponent of the Structure-from-Motion pipeline can quickly become a bottle-neck. H
ere we present a novel formulation for solving bundle adjustment in a truly dist
ributed manner using consensus based optimization methods. Our algorithm is pres
ented with a concise derivation based on proximal splitting, along with a theore
tical proof of convergence and brief discussions on complexity and implementatio
n. Experiments on a number of real image datasets convincingly demonstrates the
potential of the proposed method by outperforming the conventional bundle adjust
ment formulation by orders of magnitude.
****

Globally Optimal Manhattan Frame Estimation in Real-Time
Kyungdon Joo, Tae-Hyun Oh, Junsik Kim, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1763-1771
Given a set of surface normals, we pose a Manhattan Frame (MF) estimation problem as a consensus set maximization that maximizes the number of inliers over the rotation search space. We solve this problem through a branch-and-bound framework, which mathematically guarantees a globally optimal solution. However, the computational time of conventional branch-and-bound algorithms are intractable for real-time performance. In this paper, we propose a novel bound computation method within an efficient measurement domain for MF estimation, i.e., the extended Gaussian image (EGI). By relaxing the original problem, we can compute the bounds in real-time, while preserving global optimality. Furthermore, we quantitatively and qualitatively demonstrate the performance of the proposed method for synthetic and real-world data. We also show the versatility of our approach through two applications: extension to multiple MF estimation and video stabilization.
****

Mirror Surface Reconstruction Under an Uncalibrated Camera
Kai Han, Kwan-Yee K. Wong, Dirk Schnieders, Miaomiao Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1772-1780
This paper addresses the problem of mirror surface reconstruction, and a solution based on observing the reflections of a moving reference plane on the mirror surface is proposed. Unlike previous approaches which require tedious work to calibrate the camera, our method can recover both the camera intrinsics and extrinsics together with the mirror surface from reflections of the reference plane under at least three unknown distinct poses. Our previous work has demonstrated that 3D poses of the reference plane can be registered in a common coordinate system using reflection correspondences established across images. This leads to a bunch of registered 3D lines formed from the reflection correspondences. Given these lines, we first derive an analytical solution to recover the camera projection matrix through estimating the line projection matrix. We then optimize the camera projection matrix by minimizing reprojection errors computed based on a cross-ratio formulation. The mirror surface is finally reconstructed based on the optimized cross-ratio constraint. Experimental results on both synthetic and real data are presented, which demonstrate the feasibility and accuracy of our method.
****

A Hole Filling Approach Based on Background Reconstruction for View Synthesis in 3D Video
Guibo Luo, Yuesheng Zhu, Zhaotian Li, Liming Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1781-1789
The depth image based rendering (DIBR) plays a key role in 3D video synthesis, by which other virtual views can be generated from a 2D video and its depth map. However, in the synthesis process, the background occluded by the foreground objects might be exposed in the new view, resulting in some holes in the synthetized video. In this paper, a hole filling approach based on background reconstruction is proposed, in which the temporal correlation information in both the 2D video and its corresponding depth map are exploited to construct a background video. To construct a clean background video, the foreground objects are detected and removed. Also motion compensation is applied to make the background reconstruction model suitable for moving camera scenario. Each frame is projected to the current plane where a modified Gaussian mixture model is performed. The constructed background video is used to eliminate the holes in the synthetized video. Our experimental results have indicated that the proposed approach has better quality of the synthetized 3D video compared with the other methods.
****

A Direct Least-Squares Solution to the PnP Problem With Unknown Focal Length
Yinqiang Zheng, Laurent Kneip; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1790-1798
In this work, we propose a direct least-squares solution to the perspective-(n)-

point (P(n)P) pose estimation problem of a partially calibrated camera, whose in trinsic parameters except the focal length are known. The basic idea is to const ruct a proper objective function with respect to the target variables and extrac t all its stationary points so as to find the global minimum. The advantages of our proposed solution over existing ones are that (i) the objective function is directly built upon the imaging equation, such that all the 3D-to-2D corresponde nces are treated with balance, and that (ii) the proposed solution is noniterati ve, in the sense that the stationary points are retrieved by means of standard e igenvalue factorization and the common iterative refinement step is not needed. In addition, the proposed solution has (O(n)) complexity, and can be used to han dle both planar and nonplanar 3D points. Experimental results have shown that th e proposed solution is much more accurate than the existing state-of-the-art sol utions, and is even comparable to the maximum likelihood estimation by minimizin g the reprojection error.

****

Efficient Intersection of Three Quadrics and Applications in Computer Vision
Zuzana Kukelova, Jan Heller, Andrew Fitzgibbon; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1799-1808

In this paper, we present a new algorithm for finding all intersections of three quadrics. The proposed method is algebraic in nature and it is considerably mor e efficient than the Groebner basis and resultant-based solutions previously use d in computer vision applications. We identify several computer vision problems that are formulated and solved as systems of three quadratic equations and for w hich our algorithm readily delivers considerably faster results. Also, we propos e new formulations of three important vision problems: absolute camera pose with unknown focal length, generalized pose-and-scale, and hand-eye calibration with known translation. These new formulations allow our algorithm to significantly outperform the state-of-the-art in speed.

****

Using Spatial Order to Boost the Elimination of Incorrect Feature Matches
Lior Talker, Yael Moses, Ilan Shimshoni; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1809-1817

Correctly matching feature points in a pair of images is an important preprocess ing step for many computer vision applications. In this paper we propose an effi cient method for estimating the number of correct matches without explicitly com puting them. In addition, our method estimates the region of overlap between the images. To this end, we propose to analyze the set of matches using the spatial order of the features, as projected to the x-axis of the image. The set of feat ures in each image is thus represented by a sequence. This reduces the analysis of the matching problem to the analysis of the permutation between the sequences . Using the Kendall distance metric between permutations and natural assumptions on the distribution of the correct and incorrect matches, we show how to estima te the above-mentioned values. We demonstrate the usefulness of our method in tw o applications: (i) a new halting condition for RANSAC based epipolar geometry e stimation methods that considerably reduce the running time, and (ii) discarding spatially unrelated image pairs in the Structure-from-Motion pipeline. Furtherm ore, our analysis allows to compute the probability that a given match is correc t based on the estimated number of correct matches and the rank of the features within the sequences. Our experiments on a large number of synthetic and real da ta demonstrate the effectiveness of our method. For example, the running time of the image matching stage in the Structure-from-Motion pipeline may be reduced b y about 99% while preserving about 80% of the correctly matched feature points.

****

A Probabilistic Framework for Color-Based Point Set Registration
Martin Danelljan, Giulia Meneghetti, Fahad Shahbaz Khan, Michael Felsberg; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1818-1826

In recent years, sensors capable of measuring both color and depth information h ave become increasingly popular. Despite the abundance of colored point set data , state-of-the-art probabilistic registration techniques ignore the available co

lor information. In this paper, we propose a probabilistic point set registratio n framework that exploits available color information associated with the points . Our method is based on a model of the joint distribution of 3D-point observati ons and their color information. The proposed model captures discriminative colo r information, while being computationally efficient. We derive an EM algorithm for jointly estimating the model parameters and the relative transformations. C omprehensive experiments are performed on the Stanford Lounge dataset, captured by an RGB-D camera, and two point sets captured by a Lidar sensor. Our results d emonstrate a significant gain in robustness and accuracy when incorporating colo r information. On the Stanford Lounge dataset, our approach achieves a relative reduction of the failure rate by 78% compared to the baseline. Furthermore, our proposed model outperforms standard strategies for combining color and 3D-point information, leading to state-of-the-art results.
****

Blind Image Deconvolution by Automatic Gradient Activation
Dong Gong, Mingkui Tan, Yanning Zhang, Anton van den Hengel, Qinfeng Shi; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1827-1836
Blind image deconvolution is an ill-posed inverse problem which is often address ed through the application of appropriate prior. Although some priors are inform ative in general, many images do not strictly conform to this, leading to degrad ed performance in the kernel estimation. More critically, real images may be con taminated by nonuniform noise such as saturation and outliers.Methods for removi ng specific image areas based on some priors have been proposed, but they operat e either manually or by defining fixed criteria. We show here that a subset of t he image gradients are adequate to estimate the blur kernel robustly, no matter the gradient image is sparse or not. We thus introduce a gradient activation met hod to automatically select a subset of gradients of the latent image in a cutti ng-plane-based optimization scheme for kernel estimation. No extra assumption is used in our model, which greatly improves the accuracy and flexibility. More im portantly, the proposed method affords great convenience for handling noise and outliers. Experiments on both synthetic data and real-world images demonstrate t he effectiveness and robustness of the proposed method in comparison with the st ate-of-the-art methods.
****

PSyCo: Manifold Span Reduction for Super Resolution
Eduardo Perez-Pellitero, Jordi Salvador, Javier Ruiz-Hidalgo, Bodo Rosenhahn; Pr oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP R), 2016, pp. 1837-1845
The main challenge in Super Resolution (SR) is to discover the mapping between t he low- and high-resolution manifolds of image patches, a complex ill-posed prob lem which has recently been addressed through piecewise linear regression with p romising results. In this paper we present a novel regression-based SR algorithm that benefits from an extended knowledge of the structure of both manifolds. We propose a transform that collapses the 16 variations induced from the dihedral group of transforms (i.e. rotations, vertical and horizontal reflections) and an tipodality (i.e. diametrically opposed points in the unitary sphere) into a sing le primitive. The key idea of our transform is to study the different dihedral e lements as a group of symmetries within the high-dimensional manifold. We obtain the respective set of mirror-symmetry axes by means of a frequency analysis of the dihedral elements, and we use them to collapse the redundant variability thr ough a modified symmetry distance. The experimental validation of our algorithm shows the effectiveness of our approach, which obtains competitive quality with a dictionary of as little as 32 atoms (reducing other methods' dictionaries by a t least a factor of 32) and further pushing the state-of-the-art with a 1024 ato ms dictionary.
****

Parametric Object Motion From Blur
Jochen Gast, Anita Sellent, Stefan Roth; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1846-1854

Motion blur can adversely affect a number of vision tasks, hence it is generally considered a nuisance. We instead treat motion blur as a useful signal that allows to compute the motion of objects from a single image. Drawing on the success of joint segmentation and parametric motion models in the context of optical flow estimation, we propose a parametric object motion model combined with a segmentation mask to exploit localized, non-uniform motion blur. Our parametric image formation model is differentiable w.r.t. the motion parameters, which enables us to generalize marginal-likelihood techniques from uniform blind deblurring to localized, non-uniform blur. A two-stage pipeline, first in derivative space and then in image space, allows to estimate both parametric object motion as well as a motion segmentation from a single image alone. Our experiments demonstrate its ability to cope with very challenging cases of object motion blur.
****

Image Deblurring Using Smartphone Inertial Sensors
Zhe Hu, Lu Yuan, Stephen Lin, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1855-1864
Removing image blur caused by camera shake is an ill-posed problem, as both the latent image and the point spread function (PSF) are unknown. A recent approach to address this problem is to record camera motion through inertial sensors, i.e., gyroscopes and accelerometers, and then reconstruct spatially-variant PSFs from these readings. While this approach has been effective for high-quality inertial sensors, it has been infeasible for the inertial sensors in smartphones, which are of relatively low quality and present a number of challenging issues, including varying sensor parameters, high sensor noise, and calibration error. In this paper, we identify the issues that plague smartphone inertial sensors and propose a solution that successfully utilizes the sensor readings for image deblurring. With both the sensor data and the image itself, the proposed method is able to accurately estimate the sensor parameters online and also the spatially-variant PSFs for enhanced deblurring performance. The effectiveness of this technique is demonstrated in experiments on a popular mobile phone. With this approach, the quality of image deblurring can be appreciably raised on the most common of imaging devices.
****

Seven Ways to Improve Example-Based Single Image Super Resolution
Radu Timofte, Rasmus Rothe, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1865-1873
In this paper we present seven techniques that everybody should know to improve example-based single image super resolution (SR): 1) augmentation of data, 2) use of large dictionaries with efficient search structures, 3) cascading, 4) image self-similarities, 5) back projection refinement, 6) enhanced prediction by consistency check, and 7) context reasoning. We validate our seven techniques on standard SR benchmarks (i.e. Set5, Set14, B100) and methods (i.e. A+, SRCNN, ANR, Zeyde, Yang) and achieve substantial improvements.The techniques are widely applicable and require no changes or only minor adjustments of the SR methods. Moreover, our Improved A+ (IA) method sets new state-of-the-art results outperforming A+ by up to 0.9dB on average PSNR whilst maintaining a low time complexity.
****

Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network
Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, Zehan Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874-1883
Recently, several models based on deep neural networks have achieved great success in terms of both reconstruction accuracy and computational performance for single image super-resolution. In these methods, the low resolution (LR) input image is upscaled to the high resolution (HR) space using a single filter, commonly bicubic interpolation, before reconstruction. This means that the super-resolution (SR) operation is performed in HR space. We demonstrate that this is sub-optimal and adds computational complexity. In this paper, we present the first convolutional neural network (CNN) capable of real-time SR of 1080p videos on a sin

gle K2 GPU. To achieve this, we propose a novel CNN architecture where the featu re maps are extracted in the LR space. In addition, we introduce an efficient su b-pixel convolution layer which learns an array of upscaling filters to upscale the final LR feature maps into the HR output. By doing so, we effectively replac e the handcrafted bicubic filter in the SR pipeline with more complex upscaling filters specifically trained for each feature map, whilst also reducing the comp utational complexity of the overall SR operation. We evaluate the proposed appro ach using images and videos from publicly available datasets and show that it pe rforms significantly better (+0.15dB on Images and +0.39dB on Videos) and is an order of magnitude faster than previous CNN-based methods.
****

They Are Not Equally Reliable: Semantic Event Search Using Differentiated Concep t Classifiers
Xiaojun Chang, Yao-Liang Yu, Yi Yang, Eric P. Xing; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1884-1893
Complex event detection on unconstrained Internet videos has seen much progress in recent years. However, state-of-the-art performance degrades dramatically whe n the number of positive training exemplars falls short. Since label acquisition is costly, laborious, and time-consuming, there is a real need to consider the much more challenging semantic event search problem, where no example video is given. In this paper, we present a state-of-the-art event search system without any example videos. Relying on the key observation that events (e.g. dog show) a re usually compositions of multiple mid-level concepts (e.g. "dog," "theater," a nd "dog jumping"), we first train a skip-gram model to measure the relevance of each concept with the event of interest. The relevant concept classifiers then c ast votes on the test videos but their reliability, due to lack of labeled train ing videos, has been largely unaddressed. We propose to combine the concept clas sifiers based on a principled estimate of their accuracy on the unlabeled test v ideos. A novel warping technique is proposed to improve the performance and an e fficient highly-scalable algorithm is provided to quickly solve the resulting op timization. We conduct extensive experiments on the latest TRECVID MEDTest 2014, MEDTest 2013 and CCV datasets, and achieve state-of-the-art performances.
****

Going Deeper into First-Person Activity Recognition
Minghuang Ma, Haoqi Fan, Kris M. Kitani; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 1894-1903
We bring together ideas from recent work on feature design for egocentric action recognition under one framework by exploring the use of deep convolutional neur al networks (CNN). Recent work has shown that features such as hand appearance, object attributes, local hand motion and camera ego-motion are important for cha racterizing first-person actions. To integrate these ideas under one framework, we propose a twin stream network architecture, where one stream analyzes appeara nce information and the other stream analyzes motion information. Our appearance stream encodes prior knowledge of the egocentric paradigm by explicitly trainin g the network to segment hands and localize objects. By visualizing certain neur on activation of our network, we show that our proposed architecture naturally l earns features that capture object attributes and hand-object configurations. Ou r extensive experiments on benchmark egocentric action datasets show that our de ep architecture enables recognition rates that significantly outperform state-of -the-art techniques - an average 6.6% increase in accuracy over all datasets. Fu rthermore, by learning to recognize objects, actions and activities jointly, the performance of individual recognition tasks also increase by 30% (actions) and 14% (objects). We also include the results of extensive ablative analysis to hig hlight the importance of network design decisions.
****

Cascaded Interactional Targeting Network for Egocentric Video Analysis
Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 19 04-1913
Knowing how hands move and what object is being manipulated are two key sub-task

s for analyzing first-person (egocentric) action. However, lack of fully annotat ed hand data as well as imprecise foreground segmentation make either sub-task c hallenging. This work aims to explicitly address these two issues via introducin g a cascaded interactional targeting (i.e., infer both hand and active object re gions) deep neural network. Firstly, a novel EM-like learning framework is propo sed to train the pixel-level deep convolutional neural network (DCNN) by seamles sly integrating weakly supervised data (i.e., massive bounding box annotations) with a small set of strongly supervised data (i.e., fully annotated hand segment ation maps) to achieve state-of-the-art hand segmentation performance. Secondly, the resulting high-quality hand segmentation maps are further paired with the c orresponding motion maps and object feature maps, in order to explore the contex tual information among object, motion and hand to generate interactional foregro und regions (operated objects). The resulting interactional target maps (hand + active object) from our cascaded DCNN are further utilized to form discriminativ e action representation. Experiments show that our framework has achieved the st ate-of-the-art egocentric action recognition performance on the benchmark datase t Activities of Daily Living (ADL).

****

Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Unt rimmed Videos

Fabian Caba Heilbron, Juan Carlos Niebles, Bernard Ghanem; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1914-1923

In many large-scale video analysis scenarios, one is interested in localizing an d recognizing human activities that occur in short temporal intervals within lon g untrimmed videos. Current approaches for activity detection still struggle to handle large-scale video collections and the task remains relatively unexplored. This is in part due to the computational complexity of current action recogniti on approaches and the lack of a method that proposes fewer intervals in the vide o, where activity processing can be focused. In this paper, we introduce a propo sal method that aims to recover temporal segments containing actions in untrimme d videos. Building on techniques for learning sparse dictionaries, we introduce a learning framework to represent and retrieve activity proposals. We demonstrat e the capabilities of our method in not only producing high quality proposals bu t also in its efficiency. Finally, we show the positive impact our method has on recognition performance when it is used for action detection, while running at 10FPS.

****

Discriminative Hierarchical Rank Pooling for Activity Recognition

Basura Fernando, Peter Anderson, Marcus Hutter, Stephen Gould; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1 924-1932

We present hierarchical rank pooling, a video sequence encoding method for activ ity recognition. It consists of a network of rank pooling functions which captur es the dynamics of rich convolutional neural network features within a video seq uence. By stacking non-linear feature functions and rank pooling over one anothe r, we obtain a high capacity dynamic encoding mechanism, which is used for actio n recognition. We present a method for jointly learning the video representation and activity classifier parameters. Our method obtains state-of-the art results on three important activity recognition benchmarks: 76.7% on Hollywood2, 66.9% on HMDB51 and, 91.4% on UCF101.

****

Convolutional Two-Stream Network Fusion for Video Action Recognition

Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933-1941

Recent applications of Convolutional Neural Networks (ConvNets) for human action recognition in videos have proposed different solutions for incorporating the a ppearance and motion information. We study a number of ways of fusing ConvNet to wers both spatially and temporally in order to best take advantage of this spati o-temporal information. We make the following findings: (i) that rather than fus

ing at the softmax layer, a spatial and temporal network can be fused at a convolution layer without loss of performance, but with a substantial saving in parameters; (ii) that it is better to fuse such networks spatially at the last convolutional layer than earlier, and that additionally fusing at the class prediction layer can boost accuracy; finally (iii) that pooling of abstract convolutional features over spatiotemporal neighbourhoods further boosts performance. Based on these studies we propose a new ConvNet architecture for spatiotemporal fusion of video snippets, and evaluate its performance on standard benchmarks where this architecture achieves state-of-the-art results.
****

Learning Activity Progression in LSTMs for Activity Detection and Early Detection

Shugao Ma, Leonid Sigal, Stan Sclaroff; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1942-1950

In this work we improve training of temporal deep models to better learn activity progression for activity detection and early detection. Conventionally, when training a Recurrent Neural Network, specifically a Long Short Term Memory (LSTM) model, the training loss only considers classification error. However, we argue that the detection score of the correct activity category or the detection score margin between the correct and incorrect categories should be monotonically non-decreasing as the model observes more of the activity. We design novel ranking losses that directly penalize the model on violation of such monotonicities, which are used together with classification loss in training of LSTM models. Evaluation on ActivityNet shows significant benefits of the proposed ranking losses in both activity detection and early detection tasks.
****

VLAD3: Encoding Dynamics of Deep Features for Action Recognition

Yingwei Li, Weixin Li, Vijay Mahadevan, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1951-1960

Previous approaches to action recognition with deep features tend to process video frames only within a small temporal region, and do not model long-range dynamic information explicitly. However, such information is important for the accurate recognition of actions, especially for the discrimination of complex activities that share sub-actions, and when dealing with untrimmed videos. Here, we propose a representation, VLAD for Deep Dynamics (VLAD^3), that accounts for different levels of video dynamics. It captures short-term dynamics with deep convolutional neural network features, relying on linear dynamic systems (LDS) to model medium-range dynamics. To account for long-range inhomogeneous dynamics, a VLAD descriptor is derived for the LDS and pooled over the whole video, to arrive at the final VLAD^3 representation. An extensive evaluation was performed on Olympic Sports, UCF101 and THUMOS15, where the use of the VLAD^3 representation leads to state-of- the-art results.
****

A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection

Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, Ming Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1961-1970

We present a multi-stream bi-directional recurrent neural network for fine-grained action detection. Recently, two-stream convolutional neural networks (CNNs) trained on stacked optical flow and image frames have been successful for action recognition in videos. Our system uses a tracking algorithm to locate a bounding box around the person, which provides a frame of reference for appearance and motion and also suppresses background noise that is not within the bounding box. We train two additional streams on motion and appearance cropped to the tracked bounding box, along with full-frame streams. Our motion streams use pixel trajectories of a frame as raw features, in which the displacement values corresponding to a moving scene point are at the same spatial position across several frames. To model long-term temporal dynamics within and between actions, the multi-

stream CNN is followed by a bi-directional Long Short-Term Memory (LSTM) layer. We show that our bi-directional LSTM network utilizes about 8 seconds of the video sequence to predict an action label. We test on two action detection datasets: the MPII Cooking 2 Dataset, and a new MERL Shopping Dataset that we introduce and make available to the community with this paper. The results demonstrate that our method significantly outperforms state-of-the-art action detection methods on both datasets.
****

A Hierarchical Deep Temporal Model for Group Activity Recognition
Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1971-1980
In group activity recognition, the temporal dynamics of the whole activity can be inferred based on the dynamics of the individual people representing the activity. We build a deep model to capture these dynamics based on LSTM (long short-term memory) models. To make use of these observations, we present a 2-stage deep temporal model for the group activity recognition problem. In our model, a LSTM model is designed to represent action dynamics of individual people in a sequence and another LSTM model is designed to aggregate person-level information for whole activity understanding. We evaluate our model over two datasets: the Collective Activity Dataset and a new volleyball dataset. Experimental results demonstrate that our proposed model improves group activity recognition performance compared to baseline methods.
****

A Hierarchical Pose-Based Approach to Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets
Ivan Lillo, Juan Carlos Niebles, Alvaro Soto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1981-1990
In this paper, we introduce a new hierarchical model for human action recognition that is able to categorize complex actions performed in videos. Our model is also able to perform spatio-temporal annotation of the atomic actions that compose the overall complex action. That is, for each atomic action, the model generates temporal atomic action annotations by inferring the starting and ending times of the atomic action, as well spatial annotations by inferring the human body parts that are involved in each atomic action. Our model has three key properties: (i) it can be trained with no spatial supervision, as it is able to automatically discover the relevant body parts from temporal action annotations only; (ii) its jointly learned poselet and actionlet representation encodes the visual variability of actions with good generalization power; (iii) its mechanism for handling noisy body pose estimates make it robust to common pose estimation errors. We experimentally evaluate the performance of our method in multiple action recognition benchmarks. Our model consistently outperform baselines and state-of-the-art action recognition methods.
****

A Key Volume Mining Deep Framework for Action Recognition
Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, Yu Qiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1991-1999
Recently, deep learning approaches have demonstrated remarkable progresses for action recognition in videos. Most existing deep frameworks equally treat every volume i.e. spatial-temporal video clip, and directly assign a video label to all volumes sampled from it. However, within a video, discriminative actions may occur sparsely in a few key volumes, and most other volumes are irrelevant to the labeled action category. Training with a large proportion of irrelevant volumes will hurt performance. To address this issue, we propose a key volume mining deep framework to identify key volumes and conduct classification simultaneously. Specifically, our framework is trained end-to-end in an EM-like loop. In the forward pass, our network mines key volumes for each action class. In the backward pass, it updates network parameters with the help of these mined key volumes. In addition, we propose "Stochastic out" to handle key volumes from multi-modalities, and an effective yet simple "unsupervised key volume proposal" method for hig

h quality volume sampling. Our experiments show that action recognition performa nce can be significantly improved by mining key volumes, and our methods achieve state-of-the-art performance on UCF101 (93.1%).
****

Improved Hamming Distance Search Using Variable Length Substrings
Eng-Jon Ong, Miroslaw Bober; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2016, pp. 2000-2008
This paper addresses the problem of ultra-large-scale search in Hamming spaces. There has been considerable research on generating compact binary codes in visio n, for example for visual search tasks. However the issue of efficient searching through huge sets of binary codes remains largely unsolved. To this end, we pro pose a novel, unsupervised approach to thresholded search in Hamming space, supp orting long codes (e.g. 512-bits) with a wide-range of Hamming distance radii. O ur method is capable of working efficiently with billions of codes delivering be tween one to three orders of magnitude acceleration, as compared to prior art. T his is achieved by relaxing the equal-size constraint in the Multi-Index Hashing approach, leading to multiple hash-tables with variable length hash-keys. Based on the theoretical analysis of the retrieval probabilities of multiple hash-tab les we propose a novel search algorithm for obtaining a suitable set of hash-key lengths. The resulting retrieval mechanism is shown empirically to improve the efficiency over the state-of-the-art, across a range of datasets, bit-depths and retrieval thresholds.
****

Shortlist Selection With Residual-Aware Distance Estimator for K-Nearest Neighbo r Search
Jae-Pil Heo, Zhe Lin, Xiaohui Shen, Jonathan Brandt, Sung-eui Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2009-2017
In this paper, we introduce a novel shortlist computation algorithm for approxim ate, high-dimensional nearest neighbor search. Our method relies on a novel dist ance estimator: the residual-aware distance estimator, that accounts for the res idual distances of data points to their respective quantized centroids, and uses it for accurate shortlist computation. Furthermore, we perform the residual-aw are distance estimation with little additional memory and computational cost thr ough simple pre-computation methods for inverted index and multi-index schemes. Because it modifies the initial shortlist collection phase, our new algorithm is applicable to most inverted indexing methods that use vector quantization. We h ave tested the proposed method with the inverted index and multi-index on a dive rse set of benchmarks including up to one billion data points with varying dimen sions, and found that our method robustly improves the accuracy of shortlists (u p to 127% relatively higher) over the state-of-the-art techniques with a compara ble or even faster computational cost.
****

Supervised Quantization for Similarity Search
Xiaojuan Wang, Ting Zhang, Guo-Jun Qi, Jinhui Tang, Jingdong Wang; Proceedings o f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p p. 2018-2026
In this paper, we address the problem of searching for semantically similar imag es from a large database. We present a compact coding approach, supervised quant ization. Our approach simultaneously learns feature selection that linearly tran sforms the database points into a low-dimensional discriminative subspace, and q uantizes the data points in the transformed space. The optimization criterion is that the quantized points not only approximate the transformed points accuratel y, but also are semantically separable: the points belonging to a class lie in a cluster that is not overlapped with other clusters corresponding to other class es, which is formulated as a classification problem. The experiments on several standard datasets show the superiority of our approach over the state-of-the art supervised hashing and unsupervised quantization algorithms.
****

Efficient Large-Scale Approximate Nearest Neighbor Search on the GPU

Patrick Wieschollek, Oliver Wang, Alexander Sorkine-Hornung, Hendrik P. A. Lensch; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2027-2035

We present a new approach for efficient approximate nearest neighbor (ANN) search in high dimensional spaces, extending the idea of Product Quantization. We propose a two level product and vector quantization tree that reduces the number of vector comparisons required during tree traversal. Our approach also includes a novel highly parallelizable re-ranking method for candidate vectors by efficiently reusing already computed intermediate values. Due to its small memory footprint during traversal the method lends itself to an efficient, parallel GPU implementation. This Product Quantization Tree approach significantly outperforms recent state of the art methods for high dimensional nearest neighbor queries on standard reference datasets. Ours is the first work that demonstrates GPU performance superior to CPU performance on high dimensional, large scale ANN problems in time-critical real-world applications, like loop-closing in videos.

****

Collaborative Quantization for Cross-Modal Similarity Search

Ting Zhang, Jingdong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2036-2045

Cross-modal similarity search is a problem about designing a search system supporting querying across content modalities, e.g., using an image to search for texts or using a text to search for images. This paper presents a compact coding solution for efficient search, with a focus on the quantization approach which has already shown the superior performance over the hashing solutions in the single-modal similarity search. We propose a cross modal quantization approach, which is among the early attempts to introduce quantization into cross-modal search. The major contribution lies in jointly learning the quantizers for both modalities through aligning the quantized representations for each pair of image and text belonging to a document. In addition, our approach simultaneously learns the common space for both modalities in which quantization is conducted to enable efficient and effective search using the Euclidean distance computed in the common space with fast distance table lookup. Experimental results compared with several competitive algorithms over three benchmark datasets demonstrate that the proposed approach achieves the state-of-the-art performance.

****

Aggregating Image and Text Quantized Correlated Components

Thi Quynh Nhi Tran, Herve Le Borgne, Michel Crucianu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2046-2054

Cross-modal tasks occur naturally for multimedia content that can be described along two or more modalities like visual content and text. Such tasks require to "translate" information from one modality to another. Methods like kernelized canonical correlation analysis (KCCA) attempt to solve such tasks by finding aligned subspaces in the description spaces of different modalities. Since they favor correlations against modality-specific information, these methods have shown some success in both cross-modal and bi-modal tasks. However, we show that a direct use of the subspace alignment obtained by KCCA only leads to coarse translation abilities. To address this problem, we first put forward here a new representation method that aggregates information provided by the projections of both modalities on their aligned subspaces. We further suggest a method relying on neighborhoods in these subspaces to complete uni-modal information. Our proposal exhibits state-of-the-art results for bi-modal classification on Pascal VOC07 and for cross-modal retrieval on FlickR 8K and FlickR 30K.

****

Efficient Indexing of Billion-Scale Datasets of Deep Descriptors

Artem Babenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2055-2063

Existing billion-scale nearest neighbor search systems have mostly been compared on a single dataset of a billion of SIFT vectors, where systems based on the Inverted Multi-Index (IMI) have been performing very well, achieving state-of-the-art recall in several milliseconds. SIFT-like descriptors, however, are quickly

being replaced with descriptors based on deep neural networks (DNN) that provide better performance for many computer vision tasks. In this paper, we introduce a new dataset of one billion descriptors based on DNNs and reveal the relative inefficiency of IMI-based indexing for such descriptors compared to SIFT data. We then introduce two new indexing structures, the Non-Orthogonal Inverted Multi-Index (NO-IMI) and the Generalized Non-Orthogonal Inverted Multi-Index (GNO-IMI). We show that due to additional flexibility, the new structures are able to adapt to DNN descriptor distribution in a better way. In particular, extensive experiments on the new dataset demonstrate that these data structures provide considerably better trade-off between the speed of retrieval and recall, given similar amount of memory, as compared to the standard Inverted Multi-Index.
****

Deep Supervised Hashing for Fast Image Retrieval
Haomiao Liu, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2064-2072
In this paper, we present a new hashing method to learn compact binary codes for highly efficient image retrieval on large-scale datasets. While the complex image appearance variations still pose a great challenge to reliable retrieval, in light of the recent progress of Convolutional Neural Networks (CNNs) in learning robust image representation on various vision tasks, this paper proposes a novel Deep Supervised Hashing (DSH) method to learn compact similarity-preserving binary code for the huge body of image data. Specifically, we devise a CNN architecture that takes pairs of images (similar/dissimilar) as training inputs and encourages the output of each image to approximate discrete values (e.g. +1/-1). To this end, a loss function is elaborately designed to maximize the discriminability of the output space by encoding the supervised information from the input image pairs, and simultaneously imposing regularization on the real-valued outputs to approximate the desired discrete values. For image retrieval, new-coming query images can be easily encoded by propagating through the network and then quantizing the network outputs to binary codes representation. Extensive experiments on two large scale datasets CIFAR-10 and NUS-WIDE show the promising performance of our method compared with the state-of-the-arts.
****

Efficient Large-Scale Similarity Search Using Matrix Factorization
Ahmet Iscen, Michael Rabbat, Teddy Furon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2073-2081
We consider the image retrieval problem of finding the images in a dataset that are most similar to a query image. Our goal is to reduce the number of vector operations and memory for performing a search without sacrificing accuracy of the returned images. We adopt a group testing formulation and design the decoding architecture using either dictionary learning or eigendecomposition. The latter is a plausible option for small-to-medium sized problems with high-dimensional global image descriptors, whereas dictionary learning is applicable in large-scale scenarios. We evaluate our approach for global descriptors obtained from both SIFT and CNN features. Experiments with standard image search benchmarks, including the Yahoo100M dataset comprising 100 million images, show that our method gives comparable (and sometimes superior) accuracy compared to exhaustive search while requiring only 10% of the vector operations and memory. Moreover, for the same search complexity, our method gives significantly better accuracy compared to approaches based on dimensionality reduction or locality sensitive hashing.
****

Incremental Object Discovery in Time-Varying Image Collections
Theodora Kontogianni, Markus Mathias, Bastian Leibe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2082-2090
Abstract In this paper, we address the problem of object discovery in time-varying, large-scale image collections. A core part of our approach is a novel Limited Horizon Minimum Spanning Tree (LH-MST) structure that closely approximates the Minimum Spanning Tree at a small fraction of the latter's computational cost. Our proposed tree structure can be created in a local neighborhood of the matching graph during image retrieval and can be efficiently updated whenever the image

database is extended. We show how the LH-MST can be used within both single-link hierarchical agglomer- ative clustering and the Iconoid Shift framework for object discovery in image collections, resulting in significant efficiency gains and making both approaches capable of incremental clustering with online updates. We evaluate our approach on a dataset of 500k images from the city of Paris and compare its results to the batch version of both clustering algorithms.
****

Detecting Migrating Birds at Night
Jia-Bin Huang, Rich Caruana, Andrew Farnsworth, Steve Kelling, Narendra Ahuja; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2091-2099
Bird migration is a critical indicator of environmental health, biodiversity, and climate change. Existing techniques for monitoring bird migration are either expensive (e.g., satellite tracking), labor-intensive (e.g., moon watching), indirect and thus less accurate (e.g., weather radar), or intrusive (e.g., attaching geolocators on captured birds). In this paper, we present a vision-based system for detecting migrating birds in flight at night. Our system takes stereo videos of the night sky as inputs, detects multiple flying birds and estimates their orientations, speeds, and altitudes. The main challenge lies in detecting flying birds of unknown trajectories under high noise level due to the low-light environment. We address this problem by incorporating stereo constraints for rejecting physically implausible configurations and gathering evidence from two (or more) views. Specifically, we develop a robust stereo-based 3D line fitting algorithm for geometric verification and a deformable part response accumulation strategy for trajectory verification. We demonstrate the effectiveness of the proposed approach through quantitative evaluation of real videos of birds migrating at night collected with near-infrared cameras.
****

When Naive Bayes Nearest Neighbors Meet Convolutional Neural Networks
Ilja Kuzborskij, Fabio Maria Carlucci, Barbara Caputo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2100-2109
Since Convolutional Neural Networks (CNNs) have become the leading learning paradigm in visual recognition, Naive Bayes Nearest Neighbor (NBNN)-based classifiers have lost momentum in the community. This is because (1) such algorithms cannot use CNN activations as input features; (2) they cannot be used as final layer of CNN architectures for end-to-end training , and (3) they are generally not scalable and hence cannot handle big data. This paper proposes a framework that addresses all these issues, thus bringing back NBNNs on the map. We solve the first by extracting CNN activations from local patches at multiple scale levels, similarly to [13]. We address simultaneously the second and third by proposing a scalable version of Naive Bayes Non-linear Learning (NBNL, [7]). Results obtained using pre-trained CNNs on standard scene and domain adaptation databases show the strength of our approach, opening a new season for NBNNs.
****

Traffic-Sign Detection and Classification in the Wild
Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, Shimin Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2110-2118
Although promising results have been achieved in the areas of traffic-sign detection and classification, few works have provided simultaneous solutions to these two tasks for realistic real world images. We make two contributions to this problem. Firstly, we have created a large traffic-sign benchmark from 100000 Tencent Street View panoramas, going beyond previous benchmarks. It provides 100 000 images containing 30000 traffic-sign instances. These images cover large variations in illuminance and weather conditions. Each traffic-sign in the benchmark is annotated with a class label, its bounding box and pixel mask. We call this benchmark Tsinghua-Tencent 100K. Secondly, we demonstrate how a robust end-to-end convolutional neural network (CNN) can simultaneously detect and classify traffic-signs. Most previous CNN image processing solutions target objects that occupy a large proportion of an image, and such networks do not work well for tar

get objects occupying only a small fraction of an image like the traffic-signs h
ere. Experimental results show the robustness of our network and its superiority
 to alternatives. The benchmark, source code and the CNN model introduced in thi
s paper is publicly available.
****

Large Scale Semi-Supervised Object Detection Using Visual and Semantic Knowledge
 Transfer

Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandrea, Robert Gaizauskas, Li
ming Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2016, pp. 2119-2128

Deep CNN-based object detection systems have achieved remarkable success on seve
ral large-scale object detection benchmarks. However, training such detectors re
quires a large number of labeled bounding boxes, which are more difficult to obt
ain than image-level annotations. Previous work addresses this issue by transfor
ming image-level classifiers into object detectors. This is done by modeling the
 differences between the two on categories with both image-level and bounding bo
x annotations, and transferring this information to convert classifiers to detec
tors for categories without bounding box annotations. We improve this previous w
ork by incorporating knowledge about object similarities from visual and semanti
c domains during the transfer process. The intuition behind our proposed method
is that visually and semantically similar categories should exhibit more common
transferable properties than dissimilar categories, e.g. a better detector would
 result by transforming the differences between a dog classifier and a dog detec
tor onto the cat class, than would by transforming from the violin class. Experi
mental results on the challenging ILSVRC2013 detection dataset demonstrate that
each of our proposed object similarity based knowledge transfer methods outperfo
rms the baseline methods. We found strong evidence that visual similarity and se
mantic relatedness are complementary for the task, and when combined notably imp
rove detection, achieving state-of-the-art detection performance in a semi-super
vised setting.
****

Exploit All the Layers: Fast and Accurate CNN Object Detector With Scale Depende
nt Pooling and Cascaded Rejection Classifiers

Fan Yang, Wongun Choi, Yuanqing Lin; Proceedings of the IEEE Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2016, pp. 2129-2137

In this paper, we investigate two new strategies to detect objects accurately an
d efficiently using deep convolutional neural network: 1) scale-dependent poolin
g and 2) layer-wise cascaded rejection classifiers. The scale-dependent pooling
(SDP) improves detection accuracy by exploiting appropriate convolutional featur
es depending on the scale of candidate object proposals. The cascaded rejection
classifiers (CRC) effectively utilize convolutional features and eliminate negat
ive object proposals in a cascaded manner, which greatly speeds up the detection
 while maintaining high accuracy. In combination of the two, our method achieves
 significantly better accuracy compared to other state-of-the-arts in three chal
lenging datasets, PASCAL object detection challenge, KITTI object detection benc
hmark and newly collected Inner-city dataset, while being more efficient.
****

Dictionary Pair Classifier Driven Convolutional Neural Networks for Object Detec
tion

Keze Wang, Liang Lin, Wangmeng Zuo, Shuhang Gu, Lei Zhang; Proceedings of the IE
EE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2138-
2146

Feature representation and object category classification are two key components
 of most object detection methods. While significant improvements have been achi
eved for deep feature representation learning, traditional SVM/softmax classifie
rs remain the dominant methods for final object category classification. However
, SVM/softmax classifiers lack the capacity of explicitly exploiting the complex
 structure of deep features, as they are purely discriminative methods. The rece
ntly proposed discriminative dictionary pair learning (DPL) model involves a fid
elity term to minimize the reconstruction loss and a discrimination term to enha

nce the discriminative capability of the learned dictionary pair, and thus is ap
propriate for balancing the representation and discrimination to boost object de
tection performance. In this paper, we propose a novel object detection system b
y unifying DPL with the convolutional feature learning. Specifically, we incorpo
rate DPL as a Dictionary Pair Classifier Layer (DPCL) into the deep architecture
, and develop an end-to-end learning algorithm for optimizing the dictionary pai
rs and the neural networks simultaneously. Moreover, we design a multi-task loss
 for guiding our model to accomplish the three correlated tasks: objectness esti
mation, categoryness computation, and bounding box regression. From the extensiv
e experiments on PASCAL VOC 2007/2012 benchmarks, our approach demonstrates the
effectiveness to substantially improve the performances over the popular existin
g object detection frameworks (e.g., R-CNN [13] and FRCN [12]), and achieves new
 state-of-the-arts.
****

Monocular 3D Object Detection for Autonomous Driving
Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, Raquel Urtasun
; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
(CVPR), 2016, pp. 2147-2156
The goal of this paper is to perform 3D object detection in single monocular ima
ges in the domain of autonomous driving. Our method first aims to generate a se
t of candidate class-specific object proposals, which are then run through a sta
ndard CNN pipeline to obtain high-quality object detections. The focus of this p
aper is on proposal generation. In particular, we propose a probabilistic model
that places object candidates in 3D using a prior on ground-plane. We then score
 each candidate box projected to the image plane via several intuitive potential
s such as semantic segmentation, contextual information, size and location prior
s and typical object shape. The weights in our model are trained with S-SVM. Exp
eriments show that our object proposal generation approach significantly outperf
orms all monocular baselines, and achieves the best detection performance on the
 challenging KITTI benchmark, among the published monocular competitors.
****

How Hard Can It Be? Estimating the Difficulty of Visual Search in an Image
Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P. Papad
opoulos, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2016, pp. 2157-2166
We address the problem of estimating image difficulty defined as the human respo
nse time for solving a visual search task. We collect human annotations of image
 difficulty for the PASCAL VOC 2012 data set through a crowd-sourcing platform.
We then analyze what human interpretable image properties can have an impact on
visual search difficulty, and how accurate are those properties for predicting d
ifficulty. Next, we build a regression model based on deep features learned with
 state of the art convolutional neural networks and show better results for pred
icting the ground-truth visual search difficulty scores produced by human annota
tors. Our model is able to correctly rank about 75% image pairs according to the
ir difficulty score. We also show that our difficulty predictor generalizes well
 to new classes not seen during training. Finally, we demonstrate that our predi
cted difficulty scores are useful for weakly supervised object localization (8%
improvement) and semi-supervised object classification (1% improvement).
****

Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles
Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, Tiejun Huang; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2
167-2175
The growing explosion in the use of surveillance cameras in public security high
lights the importance of vehicle search from a large-scale image or video databa
se. However, compared with person re-identification or face recognition, vehicle
 search problem has long been neglected by researchers in vision community. This
 paper focuses on an interesting but challenging problem, vehicle re-identificat
ion (a.k.a precise vehicle search). We propose a Deep Relative Distance Learning
 (DRDL) method which exploits a two-branch deep convolutional network to project

raw vehicle images into an Euclidean space where distance can be directly used to measure the similarity of arbitrary two vehicles. To further facilitate the future research on this problem, we also present a carefully-organized large-scale image database "VehicleID", which includes multiple images of the same vehicle captured by different real-world cameras in a city. We evaluate our DRDL method on our VehicleID dataset and another recently-released vehicle model classification dataset "CompCars" in three sets of experiments: vehicle re-identification, vehicle model verification and vehicle retrieval. Experimental results show that our method can achieve promising results and outperforms several state-of-the-art approaches.

****

Eye Tracking for Everyone

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2176-2184

From scientific research to commercial applications, eye tracking is an important tool across many domains. Despite its range of applications, eye tracking has yet to become a pervasive technology. We believe that we can put the power of eye tracking in everyone's palm by building eye tracking software that works on commodity hardware such as mobile phones and tablets, without the need for additional sensors or devices. We tackle this problem by introducing GazeCapture, the first large-scale dataset for eye tracking, containing data from over 1450 people consisting of almost 2:5M frames. Using GazeCapture, we train iTracker, a convolutional neural network for eye tracking, which achieves a significant reduction in error over previous approaches while running in real time (10-15fps) on a modern mobile device. Our model achieves a prediction error of 1.71cm and 2.53cm without calibration on mobile phones and tablets respectively. With calibration, this is reduced to 1.34cm and 2.12cm. Further, we demonstrate that the features learned by iTracker generalize well to other datasets, achieving state-of-the-art results. The code, data, and models are available at http://gazecapture.csail.mit.edu.

****

Efficient Globally Optimal 2D-To-3D Deformable Shape Matching

Zorah Lahner, Emanuele Rodola, Frank R. Schmidt, Michael M. Bronstein, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2185-2193

We propose the first algorithm for non-rigid 2D-to-3D shape matching, where the input is a 2D query shape as well as a 3D target shape and the output is a continuous matching curve represented as a closed contour on the 3D shape. We cast the problem as finding the shortest circular path on the product 3-manifold of the two shapes. We prove that the optimal matching can be computed in polynomial time with a (worst-case) complexity of O(m*n^2*log(n)), where m and n denote the number of vertices on the 2D and the 3D shape respectively. Quantitative evaluation confirms that the method provides excellent results for sketch-based deformable 3D shape retrieval.

****

Ambiguity Helps: Classification With Disagreements in Crowdsourced Annotations

Viktoriia Sharmanska, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, Novi Quadrianto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2194-2202

Imagine we show an image to a person and ask her/him to decide whether the scene in the image is warm or not warm, and whether it is easy or not to spot a squirrel in the image. For exactly the same image, the answers to those questions are likely to differ from person to person. This is because the task is inherently ambiguous. Such an ambiguous, therefore challenging, task is pushing the boundary of computer vision in showing what can and can not be learned from visual data. Crowdsourcing has been invaluable for collecting annotations. This is particularly so for a task that goes beyond a clear-cut dichotomy as multiple human judgments per image are needed to reach a consensus. This paper makes conceptual and technical contributions. On the conceptual side, we define disagreements among

annotators as privileged information about the data instance. On the technical s ide, we propose a framework to incorporate annotation disagreements into the cla ssifiers. The proposed framework is simple, relatively fast, and outperforms cla ssifiers that do not take into account the disagreements, especially if tested o n high confidence annotations.
****

A Task-Oriented Approach for Cost-Sensitive Recognition
Roozbeh Mottaghi, Hannaneh Hajishirzi, Ali Farhadi; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2203-2211
With the recent progress in visual recognition, we have already started to see a surge of vision related real-world applications. These applications, unlike gen eral scene understanding, are task oriented and require specific information fro m visual data. Considering the current growth in new sensory devices, feature de signs, feature learning methods, and algorithms, the search in the space of feat ures and models becomes combinatorial. In this paper, we propose a novel cost-se nsitive task-oriented recognition method that is based on a combination of lingu istic semantics and visual cues. Our task-oriented framework is able to generali ze to unseen tasks for which there is no training data and outperforms state-of- the-art cost-based recognition baselines on our new task-based dataset.
****

Refining Architectures of Deep Convolutional Neural Networks
Sukrit Shankar, Duncan Robertson, Yani Ioannou, Antonio Criminisi, Roberto Cipol la; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio n (CVPR), 2016, pp. 2212-2220
Deep Convolutional Neural Networks (CNNs) have recently evinced immense success for various image recognition tasks. However, a question of paramount importance is somewhat unanswered in deep learning research - is the selected CNN optimal for the dataset in terms of accuracy and model size? In this paper, we intend t o answer this question and introduce a novel strategy that alters the architectu re of a given CNN for a specified dataset, to potentially enhance the original a ccuracy while possibly reducing the model size. We use two operations for archit ecture refinement, viz. stretching and symmetrical splitting. Stretching increas es the number of hidden units (nodes) in a given CNN layer, while a symmetrical split of say K between two layers separates the input and output channels into K equal groups, and connects only the corresponding input-output channel groups. Our procedure starts with a pre-trained CNN for a given dataset, and optimally decides the stretch and split factors across the network to refine the architec ture. We empirically demonstrate the necessity of the two operations. We evalu ate our approach on two natural scenes attributes datasets, SUN Attributes and CAMIT-NSAD, with architectures of GoogleNet and VGG-11, that are quite contrasti ng in their construction. We justify our choice of datasets, and show that they are interestingly distinct from each other, and together pose a challenge to our architectural refinement algorithm. Our results substantiate the usefulness of the proposed method.
****

iLab-20M: A Large-Scale Controlled Object Dataset to Investigate Deep Learning
Ali Borji, Saeed Izadi, Laurent Itti; Proceedings of the IEEE Conference on Comp uter Vision and Pattern Recognition (CVPR), 2016, pp. 2221-2230
Tolerance to image variations (e.g. translation, scale, pose, illumination, back ground) is an important desired property of any object recognition system, be it human or machine. Moving towards increasingly bigger datasets has been trending in computer vision especially with the emergence of highly popular deep learnin g models. While being very useful for learning invariance to object inter- and i ntra-class shape variability, these large-scale wild datasets are not very usefu l for learning invariance to other parameters urging researchers to resort to o ther tricks for training a model. In this work, we introduce a large-scale synth etic dataset, which is freely and publicly available, and use it to answer sever al fundamental questions regarding selectivity and invariance properties of conv olutional neural networks. Our dataset contains two parts: a) objects shot on a turntable: 15 categories, 8 rotation angles, 11 cameras on a semi-circular arch,

5 lighting conditions, 3 focus levels, variety of backgrounds (23.4 per instanc
e) generating 1320 images per instance (about 22 million images in total), and b
) scenes: in which a robotic arm takes pictures of objects on a 1:160 scale scen
e. We study: 1) invariance and selectivity of different CNN layers, 2) knowledge
 transfer from one object category to another, 3) systematic or random sampling
of images to build a train set, 4) domain adaptation from synthetic to natural s
cenes, and 5) order of knowledge delivery to CNNs. We also discuss how our analy
ses can lead the field to develop more efficient deep learning methods.
****

Recursive Recurrent Nets With Attention Modeling for OCR in the Wild
Chen-Yu Lee, Simon Osindero; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2016, pp. 2231-2239
We present recursive recurrent neural networks with attention modeling (R2AM) fo
r lexicon-free optical character recognition in natural scene images. The primar
y advantages of the proposed method are: (1) use of recursive convolutional neur
al networks (CNNs), which allow for parametrically efficient and effective image
 feature extraction; (2) an implicitly learned character-level language model, e
mbodied in a recurrent neural network which avoids the need to use N-grams; and
(3) the use of a soft-attention mechanism, allowing the model to selectively exp
loit image features in a coordinated way, and allowing for end-to-end training w
ithin a standard backpropagation framework. We validate our method with state-of
-the-art performance on challenging benchmark datasets: Street View Text, IIIT5k
, ICDAR and Synth90k.
****

Deep Decision Network for Multi-Class Image Classification
Venkatesh N. Murthy, Vivek Singh, Terrence Chen, R. Manmatha, Dorin Comaniciu; P
roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV
PR), 2016, pp. 2240-2248
In this paper, we present a novel Deep Decision Network (DDN) that provides an a
lternative approach towards building an efficient deep learning network. During
the learning phase, starting from the root network node, DDN automatically build
s a network that splits the data into disjoint clusters of classes which would b
e handled by the subsequent expert networks. This results in a tree-like structu
red network driven by the data. The proposed method provides an insight into the
 data by identifying the group of classes that are hard to classify and require
more attention when compared to others. DDN also has the ability to make early d
ecisions thus making it suitable for time-sensitive applications. We validate D
DN on two publicly available benchmark datasets: CIFAR-10 and CIFAR-100 and it y
ields state-of-the-art classification performance on both the datasets. The prop
osed algorithm has no limitations to be applied to any generic classification pr
oblems.
****

Less Is More: Zero-Shot Learning From Online Textual Documents With Noise Suppre
ssion
Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, Anton van den Hengel; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2
249-2257
Classifying a visual concept merely from its associated online textual source, s
uch as a Wikipedia article, is an attractive research topic in zero-shot learnin
g because it alleviates the burden of manually collecting semantic attributes. S
everal recent works have pursued this approach by exploring various ways of conn
ecting the visual and text domains. This paper revisits this idea by stepping fu
rther to consider one important factor: the textual representation is usually to
o noisy for the zero-shot learning application. This consideration motivates us
to design a simple-but-effective zero-shot learning method capable of suppressin
g noise in the text. More specifically, we propose an l_2,1-norm based objective
 function which can simultaneously suppress the noisy signal in the text and lea
rn a function to match the text document and visual features. We also develop an
 optimization algorithm to efficiently solve the resulting problem. By conductin
g experiments on two large datasets, we demonstrate that the proposed method sig

nificantly outperforms the competing methods which rely on online information so urces but without explicit noise suppression. We further make an in-depth analys is of the proposed method and provide insight as to what kind of information in documents is useful for zero-shot learning.
****
Fast Algorithms for Linear and Kernel SVM+
Wen Li, Dengxin Dai, Mingkui Tan, Dong Xu, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2258-2266
The SVM+ approach has shown excellent performance in visual recognition tasks for exploiting privileged information in the training data. In this paper, we propose two efficient algorithms for solving the linear and kernel SVM+, respectively. For linear SVM+, we absorb the bias term into the weight vector, and formulate a new optimization problem with simpler constraints in the dual form. Then, we develop an efficient dual coordinate descent algorithm to solve the new optimization problem. For kernel SVM+, we further apply the l2-loss, which leads to a simpler optimization problem in the dual form with only half of dual variables when compared with the dual form of the original SVM+ method. More interestingly, we show that our new dual problem can be efficiently solved by using the SMO algorithm of the one-class SVM problem. Comprehensive experiments on three datasets clearly demonstrate that our proposed algorithms achieve significant speed-up than the state-of-the-art solvers for linear and kernel SVM+.
****
Hierarchically Gated Deep Networks for Semantic Segmentation
Guo-Jun Qi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2267-2275
Semantic segmentation aims to parse the scene structure of images by annotating the labels to each pixel so that images can be segmented into different regions. While image structures usually have various scales, it is difficult to use a single scale to model the spatial contexts for all individual pixels. Multi-scale Convolutional Neural Networks (CNNs) and their variants have made striking success for modeling the global scene structure for an image. However, they are limited in labeling fine-grained local structures like pixels and patches, since spatial contexts might be blindly mixed up without appropriately customizing their scales. To address this challenge, we develop a novel paradigm of multi-scale deep network to model spatial contexts surrounding different pixels at various scales. It builds multiple layers of memory cells, learning feature representations for individual pixels at their customized scales by hierarchically absorbing relevant spatial contexts via memory gates between layers.Such Hierarchically Gated Deep Networks (HGDNs) can customize a suitable scale for each pixel, thereby delivering better performance on labeling scene structures of various scales. We conduct the experiments on two datasets, and show competitive results compared with the other multi-scale deep networks on the semantic segmentation task.
****
Deep Structured Scene Parsing by Learning With Image Descriptions
Liang Lin, Guangrun Wang, Rui Zhang, Ruimao Zhang, Xiaodan Liang, Wangmeng Zuo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2276-2284
This paper addresses the problem of structured scene parsing, i.e., parsing the input scene into a configuration including hierarchical semantic objects with their interaction relations. We propose a deep architecture consisting of two networks: i) a convolutional neural network (CNN) extracting the image representation for pixelwise object labeling and ii) a recursive neural network (RNN) discovering the hierarchical object structure and the inter-object relations. Rather than relying on elaborative annotations (e.g., manually labeled semantic maps and relations), we train our deep model in a weakly-supervised manner by leveraging the descriptive sentences of the training images. Specifically, we decompose each sentence into a semantic tree consisting of nouns and verb phrases, and facilitate these trees discovering the configurations of the training images. Once these scene configurations are determined, then the parameters of both the CNN and

RNN are updated accordingly by back propagation. The entire model training is ac
complished through an Expectation-Maximization method. Extensive experiments sug
gest that our model is capable of producing meaningful and structured scene conf
igurations and achieving more favorable scene labeling performance on PASCAL VOC
 2012 over other state-of-the-art weakly-supervised methods.
****

CNN-RNN: A Unified Framework for Multi-Label Image Classification
Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, Wei Xu; Proceedings
 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,
 pp. 2285-2294
While deep convolutional neural networks (CNNs) have shown a great success in si
ngle-label image classification, it is important to note that most real world im
ages contain multiple labels, which could correspond to different objects, scene
s, actions and attributes in an image. Traditional approaches to multi-label ima
ge classification learn independent classifiers for each category and employ ran
king or thresholding on the classification results. These techniques, although w
orking well, fail to explicitly exploit the label dependencies in an image. In t
his paper, we utilize recurrent neural networks (RNNs) to address this problem.
Combined with CNNs, the proposed CNN-RNN framework learns a joint image-label em
bedding to characterize the semantic label dependency as well as the image-label
 relevance, and it can be trained end-to-end from scratch to integrate both info
rmation in an unified framework. Experimental results on public benchmark datase
ts demonstrate that the proposed architecture achieves better performance than t
he state-of-the-art multi-label classification models.
****

Walk and Learn: Facial Attribute Representation Learning From Egocentric Video a
nd Contextual Data
Jing Wang, Yu Cheng, Rogerio Schmidt Feris; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2295-2304
The way people look in terms of facial attributes (ethnicity, hair color, facial
 hair, etc.) and the clothes or accessories they wear (sunglasses, hat, hoodies,
 etc.) is highly dependent on geo-location and weather condition, respectively.
This work explores, for the first time, the use of this contextual information,
as people with wearable cameras walk across different neighborhoods of a city, i
n order to learn a rich feature representation for facial attribute classificati
on, without the costly manual annotation required by previous methods. By tracki
ng the faces of casual walkers on more than 40 hours of egocentric video, we are
 able to cover tens of thousands of different identities and automatically extra
ct nearly 5 million pairs of images connected by or from different face tracks,
along with their weather and location context, under pose and lighting variation
s. These image pairs are then fed into a deep network that preserves similarity
of images connected by the same track, in order to capture identity-related attr
ibute features, and optimizes for location and weather prediction to capture add
itional facial attribute features. Finally, the network is fine-tuned with manua
lly annotated samples. We perform an extensive experimental analysis on wearable
 data and two standard benchmark datasets based on web images (LFWA and CelebA).
 Our method outperforms by a large margin a network trained from scratch. Moreov
er, even without using manually annotated identity labels for pre-training as in
 previous methods, our approach achieves results that are better than the state
of the art.
****

CNN-N-Gram for Handwriting Word Recognition
Arik Poznanski, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2016, pp. 2305-2314
Given an image of a handwritten word, a CNN is employed to estimate its n-gram f
requency profile, which is the set of n-grams contained in the word. Frequencies
 for unigrams, bigrams and trigrams are estimated for the entire word and for pa
rts of it. Canonical Correlation Analysis is then used to match the estimated pr
ofile to the true profiles of all words in a large dictionary. The CNN that is u
sed employs several novelties such as the use of multiple fully connected branch

es. Applied to all commonly used handwriting recognition benchmarks, our method outperforms, by a very large margin, all existing methods.
****

Synthetic Data for Text Localisation in Natural Images
Ankush Gupta, Andrea Vedaldi, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2315-2324
In this paper we introduce a new method for text detection in natural images. The method comprises two contributions: First, a fast and scalable engine to generate synthetic images of text in clutter. This engine overlays synthetic text to existing background images in a natural way, accounting for the local 3D scene geometry. Second, we use the synthetic images to train a Fully-Convolutional Regression Network (FCRN) which efficiently performs text detection and bounding-box regression at all locations and multiple scales in an image. We discuss the relation of FCRN to the recently-introduced YOLO detector, as well as other end-to-end object detection systems based on deep learning. The resulting detection network significantly out performs current methods for text detection in natural images, achieving an F-measure of 84.2% on the standard ICDAR 2013 benchmark. Furthermore, it can process 15 images per second on a GPU.
****

End-To-End People Detection in Crowded Scenes
Russell Stewart, Mykhaylo Andriluka, Andrew Y. Ng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2325-2333
Current people detectors operate either by scanning an image in a sliding window fashion or by classifying a discrete set of proposals. We propose a model that is based on decoding an image into a set of people detections. Our system takes an image as input and directly outputs a set of distinct detection hypotheses. Because we generate predictions jointly, common post-processing steps such as non-maximum suppression are unnecessary. We use a recurrent LSTM layer for sequence generation and train our model end-to-end with a new loss function that operates on sets of detections. We demonstrate the effectiveness of our approach on the challenging task of detecting people in crowded scenes
****

Real-Time Salient Object Detection With a Minimum Spanning Tree
Wei-Chih Tu, Shengfeng He, Qingxiong Yang, Shao-Yi Chien; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2334-2342
In this paper, we present a real-time salient object detection system based on the minimum spanning tree. Due to the fact that background regions are typically connected to the image boundaries, salient objects can be extracted by computing the distances to the boundaries. However, measuring the image boundary connectivity efficiently is a challenging problem. Existing methods either rely on super pixel representation to reduce the processing units or approximate the distance transform. Instead, we propose an exact and iteration free solution on a minimum spanning tree. The minimum spanning tree representation of an image inherently reveals the object geometry information in a scene. Meanwhile, it largely reduces the search space of shortest paths, resulting an efficient and high quality distance transform algorithm. We further introduce a boundary dissimilarity measure to compliment the shortage of distance transform for salient object detection. Extensive evaluations show that the proposed algorithm achieves the leading performance compared to the state-of-the-art methods in terms of efficiency and accuracy.
****

Local Background Enclosure for RGB-D Salient Object Detection
David Feng, Nick Barnes, Shaodi You, Chris McCarthy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2343-2350
Recent work in salient object detection has considered the incorporation of depth cues from RGB-D images. In most cases, depth contrast is used as the main feature. However, areas of high contrast in background regions cause false positives for such methods, as the background frequently contains regions that are highly variable in depth. Here, we propose a novel RGB-D saliency feature. Local Backg

round Enclosure (LBE) captures the spread of angular directions which are backgr ound with respect to the candidate region and the object that it is part of. We show that our feature improves over state-of-the-art RGB-D saliency approaches a s well as RGB methods on the RGBD1000 and NJUDS2000 datasets.
****

Adaptive Object Detection Using Adjacency and Zoom Prediction
Yongxi Lu, Tara Javidi, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2351-2359
State-of-the-art object detection systems rely on an accurate set of region prop osals. Several recent methods use a neural network architecture to hypothesize p romising object locations. While these approaches are computationally efficient, they rely on fixed image regions as anchors for predictions. In this paper we p ropose to use a search strategy that adaptively directs computational resources to sub-regions likely to contain objects. Compared to methods based on fixed anc hor locations, our approach naturally adapts to cases where object instances are sparse and small. Our approach is comparable in terms of accuracy to the state- of-the-art Faster R-CNN approach while using two orders of magnitude fewer ancho rs on average. Code is publicly available.
****

Semantic Channels for Fast Pedestrian Detection
Arthur Daniel Costea, Sergiu Nedevschi; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2016, pp. 2360-2368
Pedestrian detection and semantic segmentation are high potential tasks for many real-time applications. However most of the top performing approaches provide s tate of art results at high computational costs. In this work we propose a fast solution for achieving state of art results for both pedestrian detection and se mantic segmentation. As baseline for pedestrian detection we use sliding window s over cost efficient multiresolution filtered LUV+HOG channels. We use the same channels for classifying pixels into eight semantic classes. Using short range and long range multiresolution channel features we achieve more robust segmenta tion results compared to traditional codebook based approaches at much lower com putational costs. The resulting segmentations are used as additional semantic ch annels in order to achieve a more powerful pedestrian detector. To also achieve fast pedestrian detection we employ a multiscale detection scheme based on a sin gle flexible pedestrian model and a single image scale. The proposed solution pr ovides competitive results on both pedestrian detection and semantic segmentatio n benchmarks at 8 FPS on CPU and at 15 FPS on GPU, being the fastest top perform ing approach.
****

G-CNN: An Iterative Grid Based Object Detector
Mahyar Najibi, Mohammad Rastegari, Larry S. Davis; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2369-2377
We introduce G-CNN, an object detection technique based on CNNs which works with out proposal algorithms. G-CNN starts with a multi-scale grid of fixed bounding boxes. We train a regressor to move and scale elements of the grid towards objec ts iteratively. G-CNN models the problem of object detection as finding a path f rom a fixed grid to boxes tightly surrounding the objects. G-CNN with around 180 boxes in a multi-scale grid performs comparably to Fast R-CNN which uses around 2K bounding boxes generated with a proposal technique. This strategy makes dete ction faster by removing the object proposal stage as well as reducing the numbe r of boxes to be processed.
****

Recurrent Face Aging
Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2378-2386
Modeling the aging process of human face is important for cross-age face verific ation and recognition. In this paper, we introduce a recurrent face aging (RFA) framework based on a recurrent neural network which can identify the ages of peo ple from 0 to 80. Due to the lack of labeled face data of the same person captur

ed in a long range of ages, traditional face aging models usually split the ages into discrete groups and learn a one-step face feature transformation for each pair of adjacent age groups. However, those methods neglect the in-between evolving states between the adjacent age groups and the synthesized faces often suffer from severe ghosting artifacts. Since human face aging is a smooth progression, it is more appropriate to age the face by going through smooth transition states. In this way, the ghosting artifacts can be effectively eliminated and the intermediate aged faces between two discrete age groups can also be obtained. Towards this target, we employ a two-layer gated recurrent unit as the basic recurrent module whose bottom layer encodes a young face to a latent representation and the top layer decodes the representation to a corresponding older face. The experimental results demonstrate our proposed RFA provides better aging faces over other state-of-the-art age progression methods.
****

Face2Face: Real-Time Face Capture and Reenactment of RGB Videos
Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, Matthias Niessner; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2387-2395
We present a novel approach for real-time facial reenactment of a monocular target video sequence (e.g., Youtube video). The source sequence is also a monocular video stream, captured live with a commodity webcam. Our goal is to animate the facial expressions of the target video by a source actor and re-render the manipulated output video in a photo-realistic fashion. To this end, we first address the under-constrained problem of facial identity recovery from monocular video by non-rigid model-based bundling. At run time, we track facial expressions of both source and target video using a dense photometric consistency measure. Reenactment is then achieved by fast and efficient deformation transfer between source and target. The mouth interior that best matches the re-targeted expression is retrieved from the target sequence and warped to produce an accurate fit. Finally, we convincingly re-render the synthesized target face on top of the corresponding video stream such that it seamlessly blends with the real-world illumination. We demonstrate our method in a live setup, where Youtube videos are reenacted in real time.
****

Self-Adaptive Matrix Completion for Heart Rate Estimation From Face Videos Under Realistic Conditions
Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2396-2404
Recent studies in computer vision have shown that, while practically invisible to a human observer, skin color changes due to blood flow can be captured on face videos and, surprisingly, be used to estimate the heart rate (HR). While considerable progress has been made in the last few years, still many issues remain open. In particular, state-of-the-art approaches are not robust enough to operate in natural conditions (e.g. in case of spontaneous movements, facial expressions, or illumination changes). Opposite to previous approaches that estimate the HR by processing all the skin pixels inside a fixed region of interest, we introduce a strategy to dynamically select face regions useful for robust HR estimation. Our approach, inspired by recent advances on matrix completion theory, allows us to predict the HR while simultaneously discover the best regions of the face to be used for estimation. Thorough experimental evaluation conducted on public benchmarks suggests that the proposed approach significantly outperforms state-of-the-art HR estimation methods in naturalistic conditions.
****

Visually Indicated Sounds
Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2405-2413
Objects make distinctive sounds when they are hit or scratched. These sounds reveal aspects of an object's material properties, as well as the actions that prod

uced them. In this paper, we propose the task of predicting what sound an object makes when struck as a way of studying physical interactions within a visual scene. We present an algorithm that synthesizes sound from silent videos of people hitting and scratching objects with a drumstick. This algorithm uses a recurrent neural network to predict sound features from videos and then produces a waveform from these features with an example-based synthesis procedure. We show that the sounds predicted by our model are realistic enough to fool participants in a "real or fake" psychophysical experiment, and that they convey significant information about material properties and physical interactions.

****

Image Style Transfer Using Convolutional Neural Networks

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423

Rendering the semantic content of an image in different styles is a difficult image processing task. Arguably, a major limiting factor for previous approaches has been the lack of image representations that explicitly represent semantic information and, thus, allow to separate image content from style. Here we use image representations derived from Convolutional Neural Networks optimised for object recognition, which make high level image information explicit. We introduce A Neural Algorithm of Artistic Style that can separate and recombine the image content and style of natural images. The algorithm allows us to produce new images of high perceptual quality that combine the content of an arbitrary photograph with the appearance of numerous well-known artworks. Our results provide new insights into the deep image representations learned by Convolutional Neural Networks and demonstrate their potential for high level image synthesis and manipulation.

****

Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification

Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, Joel H. Saltz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2424-2433

Convolutional Neural Networks (CNN) are state-of-the-art models for many image classification tasks. However, to recognize cancer subtypes automatically, training a CNN on gigapixel resolution Whole Slide Tissue Images (WSI) is currently computationally impossible. The differentiation of cancer subtypes is based on cellular-level visual features observed on image patch scale. Therefore, we argue that in this situation, training a patch-level classifier on image patches will perform better than or similar to an image-level classifier. The challenge becomes how to intelligently combine patch-level classification results and model the fact that not all patches will be discriminative. We propose to train a decision fusion model to aggregate patch-level predictions given by patch-level CNNs, which to the best of our knowledge has not been shown before. Furthermore, we formulate a novel Expectation-Maximization (EM) based method that automatically locates discriminative patches robustly by utilizing the spatial relationships of patches. We apply our method to the classification of glioma and non-small-cell lung carcinoma cases into subtypes. The classification accuracy of our method is similar to the inter-observer agreement between pathologists. Although it is impossible to train CNNs on WSIs, we experimentally demonstrate using a comparable non-cancer dataset of smaller images that a patch-based CNN can outperform an image-based CNN.

****

Hedgehog Shape Priors for Multi-Object Segmentation

Hossam Isack, Olga Veksler, Milan Sonka, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2434-2442

Star-convexity prior is popular for interactive single object segmentation due to its simplicity and amenability to binary graph cut optimization. We propose a more general multi-object segmentation approach. Moreover, each object can be constrained by a more descriptive shape prior, "hedgehog". Each hedgehog shape has its surface normals locally constrained by an arbitrary given vector field, e.g

. gradient of the user-scribble distance transform. In contrast to star-convexit y, the tightness of our normal constraint can be changed giving better control o ver allowed shapes. For example, looser constraints, i.e. wider cones of allowed normals, give more relaxed hedgehog shapes. On the other hand, the tightest con straint enforces skeleton consistency with the scribbles. In general, hedgehog s hapes are more descriptive than a star, which is only a special case correspondi ng to a radial vector field and weakest tightness. Our approach has significantl y more applications than standard single star-convex segmentation, e.g. in medic al data we can separate multiple non-star organs with similar appearances and we ak edges. Optimization is done by our modified a-expansion moves shown to be sub modular for multi-hedgehog shapes.

****

Latent Variable Graphical Model Selection Using Harmonic Analysis: Applications to the Human Connectome Project (HCP)

Won Hwa Kim, Hyunwoo J. Kim, Nagesh Adluru, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2443-2451

A major goal of imaging studies such as the (ongoing) Human Connectome Project (HCP) is to characterize the structural network map of the human brain and identi fy its associations with covariates such as genotype, risk factors, and so on th at correspond to an individual. But the set of image derived measures and the se t of covariates are both large, so we must first estimate a 'parsimonious' set o f relations between the measurements. For instance, a Gaussian graphical model w ill show conditional independences between the random variables, which can then be used to setup specific hypothesis based analyses downstream. But most such da ta involve a large list of 'latent' variables that remain unobserved, yet affect the 'observed' variables sustantially. Accounting for such latent variables fal ls outside the scope of standard inverse covariance matrix estimation, and is ta ckled via highly specialized optimization methods. This paper offers a unique ha rmonic analysis view of this problem. By casting the estimation of the precision matrix in terms of a composition of low-frequency latent variables and high-fre quency sparse terms, we show how the problem can be formulated using a new wavel et-type expansion in non-Euclidean spaces. Our formalization poses the estimatio n problem entirely in the frequency space and shows how it can be solved by a si mple sub-gradient scheme (involving a single variable). We provide a compelling set of scientific results on 500 scans from the recently released HCP data whe re our algorithm recovers highly interpretable and sparse conditional dependenci es between brain connectivity pathways and well-known covariates.

****

Simultaneous Estimation of Near IR BRDF and Fine-Scale Surface Geometry

Gyeongmin Choe, Srinivasa G. Narasimhan, In So Kweon; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2452-2460

Near-Infrared (NIR) images of most materials exhibit less texture or albedo vari ations making them beneficial for vision tasks such as intrinsic image decomposi tion and structured light depth estimation. Understanding the reflectance proper ties (BRDF) of materials in the NIR wavelength range can be further useful for m any photometric methods including shape from shading and inverse rendering. Howe ver, even with less albedo variation, many materials e.g. fabrics, leaves, etc. exhibit complex fine-scale surface detail making it hard to accurately estimate BRDF. In this paper, we present an approach to simultaneously estimate NIR BRDF and fine-scale surface details by imaging materials under different IR lighting and viewing directions. This is achieved by an iterative scheme that alternately estimates surface detail and NIR BRDF of materials. Our setup does not require complicated gantries or calibration and we present the first NIR dataset of 100 materials including a variety of fabrics (knits, weaves, cotton, satin, leather) , and organic (skin, leaves, jute, trunk, fur) and inorganic materials (plastic, concrete, carpet). The NIR BRDFs measured from material samples are used with a shape-from-shading algorithm to demonstrate fine-scale reconstruction of object s from a single NIR image.

****

Do It Yourself Hyperspectral Imaging With Everyday Digital Cameras
Seoung Wug Oh, Michael S. Brown, Marc Pollefeys, Seon Joo Kim; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2 461-2469
Capturing hyperspectral images requires expensive and specialized hardware that is not readily accessible to most users. Digital cameras, on the other hand, are significantly cheaper in comparison and can be easily purchased and used. In th is paper, we present a framework for reconstructing hyperspectral images by usin g multiple consumer-level digital cameras. Our approach works by exploiting the different spectral sensitivities of different camera sensors. In particular, due to the differences in spectral sensitivities of the cameras, different cameras yield different RGB measurements for the same spectral signal. We introduce an algorithm that is able to combine and convert these different RGB measurements i nto a single hyperspectral image for both indoor and outdoor scenes. This camera -based approach allows hyperspectral imaging at a fraction of the cost of most e xisting hyperspectral hardware. We validate the accuracy of our reconstruction a gainst ground truth hyperspectral images (using both synthetic and real cases) a nd show its usage on relighting applications.
****
Automatic Content-Aware Color and Tone Stylization
Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, In So Kweon; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 6, pp. 2470-2478
We introduce a new technique that automatically generates diverse, visually comp elling stylizations for a photograph in an unsupervised manner. We achieve this by learning style ranking for a given input using a large photo collection and s electing a diverse subset of matching styles for final style transfer. We also p ropose an improved technique that transfers the global color and tone of the cho sen exemplars to the input photograph while avoiding the common visual artifacts produced by the existing style transfer methods. Together, our style selection and transfer techniques produce compelling, artifact-free results on a wide rang e of input photographs, and a user study shows that our results are preferred ov er other techniques.
****
Combining Markov Random Fields and Convolutional Neural Networks for Image Synth esis
Chuan Li, Michael Wand; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2016, pp. 2479-2486
This paper studies a combination of generative Markov random field (MRF) models and discriminatively trained deep convolutional neural networks (dCNNs) for synt hesizing 2D images. The generative MRF acts on higher-levels of a dCNN feature p yramid, controling the image layout at an abstract level. We apply the method to both photographic and non-photo-realistic (artwork) synthesis tasks. The MRF re gularizer prevents over-excitation artifacts and reduces implausible feature mix tures common to previous dCNN inversion approaches, permitting synthezing photog raphic content with increased visual plausibility. Unlike standard MRF-based tex ture synthesis, the combined system can both match and adapt local features with considerable variability, yielding results far out of reach of classic generati ve MRF methods.
****
DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation
Hao Chen, Xiaojuan Qi, Lequan Yu, Pheng-Ann Heng; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2487-2496
The morphology of glands has been used routinely by pathologists to assess the m alignancy degree of adenocarcinomas. Accurate segmentation of glands from histol ogy images is a crucial step to obtain reliable morphological statistics for qua ntitative diagnosis. In this paper, we proposed an efficient deep contour-aware network (DCAN) to solve this challenging problem under a unified multi-task lear ning framework. In the proposed network, multi-level contextual features from th e hierarchical architecture are explored with auxiliary supervision for accurate

gland segmentation. When incorporated with multi-task regularization during the training, the discriminative capability of intermediate features can be further improved. Moreover, our network can not only output accurate probability maps of glands, but also depict clear contours simultaneously for separating clustered objects, which further boosts the gland segmentation performance. This unified framework can be efficient when applied to large-scale histopathological data without resorting to additional steps to generate contours based on low-level cues for post-separating. Our method won the 2015 MICCAI Gland Segmentation Challenge out of 13 competitive teams, surpassing all the other methods by a significant margin.

****

Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, Ronald M. Summers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2497-2506

Despite the recent advances in automatically describing image contents, their applications have been mostly limited to image caption datasets containing natural images (e.g., Flickr 30k, MSCOCO). In this paper, we present a deep learning model to efficiently detect a disease from an image and annotate its contexts (e.g., location, severity and the affected organs). We employ a publicly available radiology dataset of chest x-rays and their reports, and use its image annotations to mine disease names to train convolutional neural networks (CNNs). In doing so, we adopt various regularization techniques to circumvent the large normal-vs-diseased cases bias. Recurrent neural networks (RNNs) are then trained to describe the contexts of a detected disease, based on the deep CNN features. Moreover, we introduce a novel approach to use the weights of the already trained pair of CNN/RNN on the domain-specific image/text dataset, to infer the joint image/text contexts for composite image labeling. Significantly improved image annotation results are demonstrated using the recurrent neural cascade model by taking the joint image/text contexts into account.

****

Conformal Surface Alignment With Optimal Mobius Search

Huu Le, Tat-Jun Chin, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2507-2516

Deformations of surfaces with the same intrinsic shape can often be described accurately by a conformal model. A major focus of computational conformal geometry is the estimation of the conformal mapping that aligns a given pair of object surfaces. The uniformization theorem en- ables this task to be acccomplished in a canonical 2D do- main, wherein the surfaces can be aligned using a Mobius transformation. Current algorithms for estimating Mobius transformations, however, often cannot provide satisfactory alignment or are computationally too costly. This paper in- troduces a novel globally optimal algorithm for estimating Mobius transformations to align surfaces that are topologi- cal discs. Unlike previous methods, the proposed algorithm deterministically calculates the best transformation, with- out requiring good initializations. Further, our algorithm is also much faster than previous techniques in practice. We demonstrate the efficacy of our algorithm on data commonly used in computational conformal geometry.

****

Coupled Harmonic Bases for Longitudinal Characterization of Brain Networks

Seong Jae Hwang, Nagesh Adluru, Maxwell D. Collins, Sathya N. Ravi, Barbara B. Bendlin, Sterling C. Johnson, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2517-2525

There is a great deal of interest in using large scale brain imaging studies to understand how brain connectivity evolves over time for an individual and how it varies over different levels/quantiles of cognitive function. To do so, one typically performs so-called tractography procedures on diffusion MR brain images and derives measures of brain connectivity expressed as graphs. The nodes correspond to distinct brain regions and the edges encode the strength of the connection. The scientific interest is in characterizing the evolution of these graphs ov

er time or from healthy individuals to diseased. We pose this important question in terms of the Laplacian of the connectivity graphs derived from various longitudinal or disease time points - quantifying its progression is then expressed in terms of coupling the harmonic bases of a full set of Laplacians. We derive a coupled system of generalized eigenvalue problems (and corresponding numerical optimization schemes) whose solution helps characterize the full life cycle of brain connectivity evolution in a given dataset. Finally, we show a set of results on a diffusion MR imaging dataset of middle aged people at risk for Alzheimer's disease (AD), who are cognitively healthy. In such asymptomatic adults, we find that a framework for characterizing brain connectivity evolution provides the ability to predict cognitive scores for individual subjects, and for estimating the progression of participant's brain connectivity into the future.
****

Automating Carotid Intima-Media Thickness Video Interpretation With Convolutional Neural Networks

Jae Shin, Nima Tajbakhsh, R. Todd Hurst, Christopher B. Kendall, Jianming Liang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2526-2535

Cardiovascular disease (CVD) is the leading cause of mortality yet largely preventable, but the key to prevention is to identify at risk individuals before adverse events. For predicting individual CVD risk, carotid intima-media thickness (CIMT), a noninvasive ultrasound method, has proven to be valuable, offering several advantages over CT coronary artery calcium score. However, each CIMT examination includes several ultrasound videos, and interpreting each of these CIMT videos involves three operations: (1) select three end-diastolic ultrasound frames (EUF) in the video, (2) localize a region of interest (ROI) in each selected frame, and (3) trace the lumen-intima interface and the media-adventitia interface in each ROI to measure CIMT. These operations are tedious, laborious, and time consuming, a serious limitation that hinders the widespread utilization of CIMT in clinical practice. To overcome this limitation, this paper presents a new system to automate CIMT video interpretation. Our extensive experiments demonstrate that the suggested system significantly outperforms the state-of-the-art methods. The superior performance is attributable to our unified framework based on convolutional neural networks (CNNs) coupled with our informative image representation and effective post-processing of the CNN outputs, which are uniquely designed for each of the above three operations.
****

Context Encoders: Feature Learning by Inpainting

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2536-2544

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders -- a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.
****

Comparative Deep Learning of Hybrid Representations for Image Recommendations

Chenyi Lei, Dong Liu, Weiping Li, Zheng-Jun Zha, Houqiang Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 25

45-2553
In many image-related tasks, learning expressive and discriminative representations of images is essential, and deep learning has been studied for automating the learning of such representations. Some user-centric tasks, such as image recommendations, call for effective representations of not only images but also preferences and intents of users over images. Such representations are termed hybrid and addressed via a deep learning approach in this paper. We design a dual-net deep network, in which the two sub-networks map input images and preferences of users into a same latent semantic space, and then the distances between images and users in the latent space are calculated to make decisions. We further propose a comparative deep learning (CDL) method to train the deep network, using a pair of images compared against one user to learn the pattern of their relative distances. The CDL embraces much more training data than naive deep learning, and thus achieves superior performance than the latter, with no cost of increasing network complexity. Experimental results with real-world data sets for image recommendations have shown the proposed dual-net network and CDL greatly outperform other state-of-the-art image recommendation solutions.
****

Fast ConvNets Using Group-Wise Brain Damage
Vadim Lebedev, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2554-2564
We revisit the idea of brain damage, i.e. the pruning of the coefficients of a neural network, and suggest how brain damage can be modified and used to speedup convolutional layers in ConvNets. The approach uses the fact that many efficient implementations reduce generalized convolutions to matrix multiplications. The suggested brain damage process prunes the convolutional kernel tensor in a group-wise fashion. After such pruning, convolutions can be reduced to multiplications of thinned dense matrices, which leads to speedup. We investigate different ways to add group-wise prunning to the learning process, and show that several-fold speedups of convolutional layers can be attained using group-sparsity regularizers. Our approach can adjust the shapes of the receptive fields in the convolutional layers, and even prune excessive feature maps from ConvNets, all in a data-driven way.
****

Learning to Co-Generate Object Proposals With a Deep Structured Network
Zeeshan Hayder, Xuming He, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2565-2573
Generating object proposals has become a key component of modern object detection pipelines. However, most existing methods generate the object candidates independently of each other. In this paper, we present an approach to co-generating object proposals in multiple images, thus leveraging the collective power of multiple object candidates. In particular, we introduce a deep structured network that jointly predicts the objectness scores and the bounding box locations of multiple object candidates. Our deep structured network consists of a fully-connected Conditional Random Field built on top of a set of deep Convolutional Neural Networks, which learn features to model both the individual object candidate and the similarity between multiple candidates. To train our deep structured network, we develop an end-to-end learning algorithm that, by unrolling the CRF inference procedure, lets us backpropagate the loss gradient throughout the entire structured network. We demonstrate the effectiveness of our approach on two benchmark datasets, showing significant improvement over state-of-the-art object proposal algorithms.
****

DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks
Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574-2582
State-of-the-art deep neural networks have achieved impressive results on many image classification tasks. However, these same architectures have been shown to be unstable to small, well sought, perturbations of the images. Despite the impo

rtance of this phenomenon, no effective methods have been proposed to accurately compute the robustness of state-of-the-art deep classifiers to such perturbations on large-scale datasets. In this paper, we fill this gap and propose the DeepFool algorithm to efficiently compute perturbations that fool deep networks, and thus reliably quantify the robustness of these classifiers. Extensive experimental results show that our approach outperforms recent methods in the task of computing adversarial perturbations and making classifiers more robust.
****

Blockout: Dynamic Model Selection for Hierarchical Deep Networks
Calvin Murdock, Zhen Li, Howard Zhou, Tom Duerig; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2583-2591
Most deep architectures for image classification--even those that are trained to classify a large number of diverse categories--learn shared image representations with a single model. Intuitively, however, categories that are more similar should share more information than those that are very different. While hierarchical deep networks address this problem by learning separate features for subsets of related categories, current implementations require simplified models using fixed architectures specified via heuristic clustering methods. Instead, we propose Blockout, a method for regularization and model selection that simultaneously learns both the model architecture and parameters. A generalization of Dropout, our approach gives a novel parametrization of hierarchical architectures that allows for structure learning via back-propagation. To demonstrate its utility, we evaluate Blockout on the CIFAR and ImageNet datasets, demonstrating improved classification accuracy, better regularization performance, faster training, and the clear emergence of hierarchical network structures.
****

FireCaffe: Near-Linear Acceleration of Deep Neural Network Training on Compute Clusters
Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Kurt Keutzer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2592-2600
Long training times for high-accuracy deep neural networks (DNNs) impede research into new DNN architectures and slow the development of high-accuracy DNNs. In this paper we present FireCaffe, which successfully scales deep neural network training across a cluster of GPUs. We also present a number of best practices to aid in comparing advancements in methods for scaling and accelerating the training of deep neural networks. The speed and scalability of distributed algorithms is almost always limited by the overhead of communicating between servers; DNN training is not an exception to this rule. Therefore, the key consideration here is to reduce communication overhead wherever possible, while not degrading the accuracy of the DNN models that we train. Our approach has three key pillars. First, we select network hardware that achieves high bandwidth between GPU servers -- Infiniband or Cray interconnects are ideal for this. Second, we consider a number of communication algorithms, and we find that reduction trees are more efficient and scalable than the traditional parameter server approach. Third, we optionally increase the batch size to reduce the total quantity of communication during DNN training, and we identify hyperparameters that allow us to reproduce the small-batch accuracy while training with large batch sizes. When training GoogLeNet and Network-in-Network on ImageNet, we achieve a 47x and 39x speedup, respectively, when training on a cluster of 128 GPUs.
****

MDL-CW: A Multimodal Deep Learning Framework With Cross Weights
Sarah Rastegar, Mahdieh Soleymani, Hamid R. Rabiee, Seyed Mohsen Shojaee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2601-2609
Deep learning has received much attention as of the most powerful approaches for multimodal representation learning in recent years. An ideal model for multimodal data can reason about missing modalities using the available ones, and usually provides more information when multiple modalities are being considered. All the previous deep models contain separate modality-specific networks and find a s

hared representation on top of those networks. Therefore, they only consider hig h level interactions between modalities to find a joint representation for them. In this paper, we propose a multimodal deep learning framework (MDL-CW) that ex ploits the cross weights between representation of modalities, and try to gradua lly learn interactions of the modalities in a deep network manner (from low to h igh level interactions). Moreover, we theoretically show that considering these interactions provide more intra-modality information, and introduce a multi-stag e pre-training method that is based on the properties of multi-modal data. In th e proposed framework, as opposed to the existing deep methods for multi-modal da ta, we try to reconstruct the representation of each modality at a given level, with representation of other modalities in the previous layer. Extensive experim ental results show that the proposed model outperforms state-of-the-art informat ion retrieval methods for both image and text queries on the PASCAL-sentence and SUN-Attribute databases.

****

Structured Receptive Fields in CNNs

Jorn-Henrik Jacobsen, Jan van Gemert, Zhongyu Lou, Arnold W. M. Smeulders; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2610-2619

Learning powerful feature representations with CNNs is hard when training data a re limited. Pre-training is one way to overcome this, but it requires large data sets sufficiently similar to the target domain. Another option is to design prio rs into the model, which can range from tuned hyperparameters to fully engineere d representations like Scattering Networks. We combine these ideas into structur ed receptive field networks, a model which has a fixed filter basis and yet reta ins the flexibility of CNNs. This flexibility is achieved by expressing receptiv e fields in CNNs as a weighted sum over a fixed basis which is similar in spirit to Scattering Networks. The key difference is that we learn arbitrary effective filter sets from the basis rather than modeling the filters. This approach expl icitly connects classical multiscale image analysis with general CNNs. With stru ctured receptive field networks, we improve considerably over unstructured CNNs for small and medium dataset scenarios as well as over Scattering for large data sets. We validate our findings on ILSVRC2012, Cifar-10, Cifar-100 and MNIST. As a realistic small dataset example, we show state-of-the-art classification resul ts on popular 3D MRI brain-disease datasets where pre-training is difficult due to a lack of large public datasets in a similar domain.

****

First Person Action Recognition Using Deep Learned Descriptors

Suriya Singh, Chetan Arora, C. V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2620-2628

We focus on the problem of wearer's action recognition in first person a.k.a. eg ocentric videos. This problem is more challenging than third person activity rec ognition due to unavailability of wearer's pose and sharp movements in the video s caused by the natural head motion of the wearer. Carefully crafted features ba sed on hands and objects cues for the problem have been shown to be successful f or limited targeted datasets. We propose convolutional neural networks (CNNs) fo r end to end learning and classification of wearer's actions. The proposed netwo rk makes use of egocentric cues by capturing hand pose, head motion and saliency map. It is compact. It can also be trained from relatively small number of labe led egocentric videos that are available. We show that the proposed network can generalize and give state of the art performance on various disparate egocentric action datasets.

****

Recognizing Micro-Actions and Reactions From Paired Egocentric Videos

Ryo Yonetani, Kris M. Kitani, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2629-2638

We aim to understand the dynamics of social interactions between two people by r ecognizing their actions and reactions using a head-mounted camera. Our work wil l impact several first-person vision tasks that need the detailed understanding of social interactions, such as automatic video summarization of group events an

d assistive systems. To recognize micro-level actions and reactions, such as sli ght shifts in attention, subtle nodding, or small hand actions, where only subtl e body motion is apparent, we propose to use paired egocentric videos recorded b y two interacting people. We show that the first-person and second-person points -of-view features of two people, enabled by paired egocentric videos, are comple mentary and essential for reliably recognizing micro-actions and reactions. We a lso build a new dataset of dyadic (two-persons) interactions that comprises more than 1000 pairs of egocentric videos to enable systematic evaluations on the ta sk of micro-action and reaction recognition.
****

Mining 3D Key-Pose-Motifs for Action Recognition
Chunyu Wang, Yizhou Wang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2639-2647
Recognizing an action from a sequence of 3D skeletal poses is a challenging task . First, different actors may perform the same action in various performing styl es. Second, the estimated poses are sometimes inaccurate due to sensory noises. These challenges can cause large variations between instances of the same class. Third, the datasets are usually small, with only a few actors performing few re petitions of each action. Hence training complex classifiers risks over-fitting the data. We address this task by mining a set of key-pose-motifs for each actio n class. A key-pose-motif contains a set of ordered poses or action units(a sh ort sequence of poses), which are required to be close but not necessarily adjac ent in the action sequences. The representation is robust to style variations an d outlier poses. The key-pose-motifs are represented in terms of a dictionary us ing soft-quantization (probabilities) to deal with inaccuracies caused by quanti zation. We propose an efficient algorithm to mine key-pose-motifs taking into ac count these probabilities. We classify a sequence by matching it to the motifs o f each class and select the class that maximizes the matching score. This simple classifier obtains state-of-the-art performance on two benchmark datasets and o utperforms a deep network approach.
****

Predicting the Where and What of Actors and Actions Through Online Action Locali zation
Khurram Soomro, Haroon Idrees, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2648-2657
This paper proposes a novel approach to tackle the challenging problem of 'onlin e action localization' which entails predicting actions and their locations as t hey happen in a video. Typically, action localization or recognition is performe d in an offline manner where all the frames in the video are processed together and action labels are not predicted for the future. This dis-allows timely local ization of actions - an important consideration for surveillance tasks. In our a pproach, given a batch of frames from the immediate past in a video, we estimate pose and over- segment the current frame into superpixels. Next, we discriminat ively train an actor foreground model on the superpixels using the pose bounding boxes. A Conditional Random Field with superpixels as nodes, and edges connecti ng spatio-temporal neighbors is used to obtain action segments. The action confi dence is predicted using dynamic programming on SVM scores obtained on short seg ments of the video, thereby capturing sequential information of the actions. The issue of visual drift is handled by updating the appearance model and pose refi nement in an online manner. Lastly, we introduce a new measure to quantify the p erformance of action prediction (i.e. online action localization), which analyze s how the prediction accuracy varies as a function of observed portion of the vi deo. Our experiments suggest that despite using only a few frames to localize ac tions at each time instant, we are able to predict the action and obtain competi tive results to state-of-the-art offline methods.
****

Actions ~ Transformations
Xiaolong Wang, Ali Farhadi, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2658-2667
What defines an action like "kicking ball"? We argue that the true meaning of an

action lies in the change or transformation an action brings to the environment. In this paper, we propose a novel representation for actions by modeling an action as a transformation which changes the state of the environment before the action happens (precondition) to the state after the action (effect). Motivated by recent advancements of video representation using deep learning, we design a Siamese network which models the action as a transformation on a high-level feature space. We show that our model gives improvements on standard action recognition datasets including UCF101 and HMDB51. More importantly, our approach is able to generalize beyond learned action categories and shows significant performance improvement on cross-category generalization on our new ACT dataset.

****

Visual Path Prediction in Complex Scenes With Crowded Moving Objects
YoungJoon Yoo, Kimin Yun, Sangdoo Yun, JongHee Hong, Hawook Jeong, Jin Young Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2668-2677
This paper proposes a novel path prediction algorithm for progressing one step further than the existing works focusing on single target path prediction. In this paper, we consider moving dynamics of co-occurring objects for path prediction in a scene that includes crowded moving objects. To solve this problem, we first suggest a two-layered probabilistic model to find major movement patterns and their co-occurrence tendency. By utilizing the unsupervised learning results from the model, we present an algorithm to find the future location of any target object. Through extensive qualitative/quantitative experiments, we show that our algorithm can find a plausible future path in complex scenes with a large number of moving objects.

****

End-To-End Learning of Action Detection From Frame Glimpses in Videos
Serena Yeung, Olga Russakovsky, Greg Mori, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2678-2687
In this work we introduce a fully end-to-end approach for action detection in videos that learns to directly predict the temporal bounds of actions. Our intuition is that the process of detecting actions is naturally one of observation and refinement: observing moments in video, and refining hypotheses about when an action is occurring. Based on this insight, we formulate our model as a recurrent neural network-based agent that interacts with a video over time. The agent observes video frames and decides both where to look next and whether to emit a prediction. Since backpropagation is not adequate in this non-differentiable setting, we use REINFORCE to learn the agent's task-specific decision policy. Our model achieves state-of-the-art results on the THUMOS'14 and ActivityNet datasets while observing only a fraction (2% or less) of the video frames.

****

Action Recognition in Video Using Sparse Coding and Relative Features
Anali Alfaro, Domingo Mery, Alvaro Soto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2688-2697
This work presents an approach to category-based action recognition in video using sparse coding techniques. The proposed approach includes two main contributions: i) A new method to handle intra-class variations by decomposing each video into a reduced set of representative atomic action acts or key-sequences, and ii) A new video descriptor, ITRA: Inter-Temporal Relational Act Descriptor, that exploits the power of comparative reasoning to capture relative similarity relations among key-sequences. In terms of the method to obtain key-sequences, we introduce a loss function that, for each video, leads to the identification of a sparse set of representative key-frames capturing both, relevant particularities arising in the input video, as well as relevant generalities arising in the complete class collection. In terms of the method to obtain the ITRA descriptor, we introduce a novel scheme to quantify relative intra and inter-class similarities among local temporal patterns arising in the videos. The resulting ITRA descriptor demonstrates to be highly effective to discriminate among action categories. As a result, the proposed approach reaches remarkable action recognition performance on several popular benchmark datasets, outperforming alternative state-of-the

-art techniques by a large margin.
****

Improving Human Action Recognition by Non-Action Classification
Yang Wang, Minh Hoai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2698-2707
In this paper we consider the task of recognizing human actions in realistic video where human actions are dominated by irrelevant factors. We first study the benefits of removing non-action video segments, which are the ones that do not portray any human action. We then learn a non-action classifier and use it to down-weight irrelevant video segments. The non-action classifier is trained using ActionThread, a dataset with shot-level annotation for the occurrence or absence of a human action. The non-action classifier can be used to identify non-action shots with high precision and subsequently used to improve the performance of action recognition systems.
****

Actionness Estimation Using Hybrid Fully Convolutional Networks
Limin Wang, Yu Qiao, Xiaoou Tang, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2708-2717
Actionness was introduced to quantify the likelihood of containing a generic action instance at a specific location. Accurate and efficient estimation of actionness is important in video analysis and may benefit other relevant tasks such as action recognition and action detection. This paper presents a new deep architecture for actionness estimation, called hybrid fully convolutional network (H-FCN), which is composed of appearance FCN (A-FCN) and motion FCN (M-FCN). These two FCNs leverage the strong capacity of deep models to estimate actionness maps from the perspectives of static appearance and dynamic motion, respectively. In addition, the fully convolutional nature of H-FCN allows it to efficiently process videos with arbitrary sizes. Experiments are conducted on the challenging datasets of Stanford40, UCF Sports, and JHMDB to verify the effectiveness of H-FCN on actionness estimation, which demonstrate that our method achieves superior performance to previous ones. Moreover, we apply the estimated actionness maps on action proposal generation and action detection. Our actionness maps advance the current state-of-the-art performance of these tasks substantially.
****

Real-Time Action Recognition With Enhanced Motion Vector CNNs
Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, Hanli Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2718-2726
The deep two-stream architecture exhibited excellent performance on video based action recognition. The most computationally expensive step in this approach comes from the calculation of optical flow which prevents it to be real-time. This paper accelerates this architecture by replacing optical flow with motion vector which can be obtained directly from compressed videos without extra calculation. However, motion vector lacks fine structures, and contains noisy and inaccurate motion patterns, leading to the evident degradation of recognition performance. Our key insight for relieving this problem is that optical flow and motion vector are inherent correlated. Transferring the knowledge learned with optical flow CNN to motion vector CNN can significantly boost the performance of the latter. Specifically, we introduce three strategies for this, initialization transfer, supervision transfer and their combination. Experimental results show that our method achieves comparable recognition performance to the state-of-the-art, while our method can process 390.7 frames per second, which is 27 times faster than the original two-stream method.
****

Laplacian Patch-Based Image Synthesis
Joo Ho Lee, Inchang Choi, Min H. Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2727-2735
Patch-based image synthesis has been enriched with global optimization on the image pyramid. Successively, the gradient-based synthesis has improved structural coherence and details. However, the gradient operator is directional and inconsi

stent and requires computing multiple operators. It also introduces a significantly heavy computational burden to solve the Poisson equation that often accompanies artifacts in non-integrable gradient fields. In this paper, we propose a patch-based synthesis using a Laplacian pyramid to improve searching correspondence with enhanced awareness of edge structures. Contrary to the gradient operators, the Laplacian pyramid has the advantage of being isotropic in detecting changes to provide more consistent performance in decomposing the base structure and the detailed localization. Furthermore, it does not require heavy computation as it employs approximation by the differences of Gaussians. We examine the potentials of the Laplacian pyramid for enhanced edge-aware correspondence search. We demonstrate the effectiveness of the Laplacian-based approach over the state-of-the-art patch-based image synthesis methods.

****

Rain Streak Removal Using Layer Priors

Yu Li, Robby T. Tan, Xiaojie Guo, Jiangbo Lu, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2736-2744

This paper addresses the problem of rain streak removal from a single image. Rain streaks impair visibility of an image and introduce undesirable interference that can severely affect the performance of computer vision algorithms. Rain streak removal can be formulated as a layer decomposition problem, with a rain streak layer superimposed on a background layer containing the true scene content. Existing decomposition methods that address this problem employ either dictionary learning methods or impose a low rank structure on the appearance of the rain streaks. While these methods can improve the overall visibility, they tend to leave too many rain streaks in the background image or over-smooth the background image. In this paper, we propose an effective method that uses simple patch-based priors for both the background and rain layers. These priors are based on Gaussian mixture models and can accommodate multiple orientations and scales of the rain streaks. This simple approach removes rain streaks better than the existing methods qualitatively and quantitatively. We overview our method and demonstrate its effectiveness over prior work on a number of examples.

****

Gradient-Domain Image Reconstruction Framework With Intensity-Range and Base-Structure Constraints

Takashi Shibata, Masayuki Tanaka, Masatoshi Okutomi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2745-2753

This paper presents a novel unified gradient-domain image reconstruction framework with intensity-range constraint and base-structure constraint. The existing method for manipulating base structures and detailed textures are classifiable into two major approaches: i) gradient-domain and ii) layer-decomposition. To generate detail-preserving and artifact-free output images, we combine the benefits of the two approaches into the proposed framework by introducing the intensity-range constraint and the base-structure constraint. To preserve details of the input image, the proposed method takes advantage of reconstructing the output image in the gradient domain, while the output intensity is guaranteed to lie within the specified intensity range, e.g. 0-to-255, by the intensity-range constraint. In addition, the reconstructed image lies close to the base structure by the base-structure constraint, which is effective for restraining artifacts. Experimental results show that the proposed framework is effective for various applications such as tone mapping, seamless image cloning, detail enhancement, and image restoration.

****

Removing Clouds and Recovering Ground Observations in Satellite Image Sequences via Temporally Contiguous Robust Matrix Completion

Jialei Wang, Peder A. Olsen, Andrew R. Conn, Aurelie C. Lozano; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2754-2763

We consider the problem of removing and replacing clouds in satellite image sequences, which has a wide range of applications in remote sensing. Our approach fi

rst detects and removes the cloud-contaminated part of the image sequences, then recovers the missing scenes from the clean parts by the proposed "TECROMAC" (TEmporally Contiguous RObust MAtrix Completion) objective. The objective function balances temporal smoothness with a low rank solution while staying close to the original observations. The matrix where the rows are pixels and columns are the days of the image has low-rank because the pixels reflect land-types such as vegetation, roads and lakes and there are relatively few of these. We provide efficient optimization algorithms for TECROMAC, so we can run on images containing millions of pixels. Empirical results on real satellite image sequences as well as simulated data demonstrate that our approach is able to recover underlying images from heavily cloud-contaminated observations.

****

D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images

Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, Thomas S. Huang ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2764-2772

In this paper, we design a Deep Dual-Domain (D3) based fast restoration model to remove artifacts of JPEG compressed images. It leverages the large learning capacity of deep networks, as well as the problem-specific expertise that was hardly incorporated in the past design of deep architectures. For the latter, we take into consideration both the prior knowledge of the JPEG compression scheme, and the successful practice of the sparsity-based dual-domain approach. We further design the One-Step Sparse Inference (1-SI) module, as an efficient and light-weighted feed-forward approximation of sparse coding. Extensive experiments verify the superiority of the proposed D3 model over several state-of-the-art methods. Specifically, our best model is capable of outperforming the latest deep model for around 1 dB in PSNR, and is 30 times faster.

****

From Bows to Arrows: Rolling Shutter Rectification of Urban Scenes

Vijay Rengarajan, Ambasamudram N. Rajagopalan, Rangarajan Aravind; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2773-2781

The rule of perspectivity that 'straight-lines-must-remain-straight' is easily inflected in CMOS cameras by distortions introduced by motion. Lines can be rendered as curves due to the row-wise exposure mechanism known as rolling shutter (RS). We solve the problem of correcting distortions arising from handheld cameras due to RS effect from a single image free from motion blur with special relevance to urban scenes. We develop a procedure to extract prominent curves from the RS image since this is essential for deciphering the varying row-wise motion. We pose an optimization problem with line desirability costs based on straightness, angle, and length, to resolve the geometric ambiguities while estimating the camera motion based on a rotation-only model assuming known camera intrinsic matrix. Finally, we rectify the RS image based on the estimated camera trajectory using inverse mapping. We show rectification results for RS images captured using mobile phone cameras. We also compare our single image method against existing video and nonblind RS rectification methods that typically require multiple images.

****

A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation

Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, Xinghao Ding; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2782-2790

We propose a weighted variational model to estimate both the reflectance and the illumination from an observed image. We show that, though it is widely adopted for ease of modeling, the log-transformed image for this task is not ideal. Based on the previous investigation of the logarithmic transformation, a new weighted variational model is proposed for better prior representation, which is imposed in the regularization terms. Different from conventional variational models, the proposed model can preserve the estimated reflectance with more details. More

over, the proposed model can suppress noise to some extent. An alternating minim ization scheme is adopted to solve the proposed model. Experimental results demo nstrate the effectiveness of the proposed model with its algorithm. Compared wit h other variational methods, the proposed method yields comparable or better res ults on both subjective and objective assessments.
****

Visualizing and Understanding Deep Texture Representations
Tsung-Yu Lin, Subhransu Maji; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2016, pp. 2791-2799
A number of recent approaches have used deep convolutional neural networks (CNNs) to build texture representations. Nevertheless, it is still unclear how these mod- els represent texture and invariances to categorical variations. This work conducts a systematic evaluation of recent CNN-based texture descriptors for rec ognition and attempts to understand the nature of invariances captured by these representations. First we show that the recently proposed bilinear CNN model [25] is an excellent generalpurpose texture descriptor and compares favorably to ot her CNN-based descriptors on various texture and scene recognition benchmarks. T he model is translationally invariant and obtains better accuracy on the ImageNe t dataset without requiring spatial jittering of data compared to corresponding models trained with spatial jittering. Based on recent work [13, 28] we propose a technique to visualize pre-images, providing a means for understanding categor ical properties that are captured by these representations. Finally, we show pre liminary results on how a unified parametric model of texture analysis and synth esis can be used for attribute-based image manipulation, e.g. to make an image m ore swirly, honeycombed, or knitted. The source code and additional visualizatio ns are available at http://vis-www.cs.umass.edu/texture.
****

Robust Kernel Estimation With Outliers Handling for Image Deblurring
Jinshan Pan, Zhouchen Lin, Zhixun Su, Ming-Hsuan Yang; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2800-2808
Estimating blur kernels from real world images is a challenging problem as the l inear image formation assumption does not hold when significant outliers, such a s saturated pixels and non-Gaussian noise, are present. While some existing non- blind deblurring algorithms can deal with outliers to a certain extent, few blin d deblurring methods are developed to well estimate the blur kernels from the bl urred images with outliers. In this paper, we present an algorithm to address th is problem by exploiting reliable edges and removing outliers in the intermediat e latent images, thereby estimating blur kernels robustly. We analyze the effect s of outliers on kernel estimation and show that most state-of-the-art blind deb lurring methods may recover delta kernels when blurred images contain significan t outliers. We propose a robust energy function which describes the properties o f outliers for the final latent image restoration. Furthermore, we show that the proposed algorithm can be applied to improve existing methods to deblur images with outliers. Extensive experiments on different kinds of challenging blurry im ages with significant amount of outliers demonstrate the proposed algorithm perf orms favorably against the state-of-the-art methods.
****

Online Collaborative Learning for Open-Vocabulary Visual Classifiers
Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, Tat-Seng Chua; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2016, pp. 2809-2817
We focus on learning open-vocabulary visual classifiers, which scale up to a lar ge portion of natural language vocabulary (e.g., over tens of thousands of class es). In particular, the training data are large-scale weakly labeled Web images since it is difficult to acquire sufficient well-labeled data at this category s cale. In this paper, we propose a novel online learning paradigm towards this c hallenging task. Different from traditional N-way independent classifiers that g enerally fail to handle the extremely sparse and inter-related labels, our class ifiers learn from continuous label embeddings discovered by collaboratively deco mposing the sparse image-label matrix. Leveraging on the structure of the propos

ed collaborative learning formulation, we develop an efficient online algorithm that can jointly learn the label embeddings and visual classifiers. The algorithm can learn over 30,000 classes of 1,000 training images within 1 second on a standard GPU. Extensively experimental results on four benchmarks demonstrate the effectiveness of our method.

****

Rethinking the Inception Architecture for Computer Vision

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826

Convolutional networks are at the core of most state-of-the-art computer vision solutions for a wide variety of tasks. Since 2014 very deep convolutional networks started to become mainstream, yielding substantial gains in various benchmarks. Although increased model size and computational cost tend to translate to immediate quality gains for most tasks (as long as enough labeled data is provided for training), computational efficiency and low parameter count are still enabling factors for various use cases such as mobile vision and big-data scenarios. Here we are exploring ways to scale up networks in ways that aim at utilizing the added computation as efficiently as possible. We benchmark our methods on the ILSVRC 2012 classification challenge validation set and demonstrate substantial gains over the state of the art via to carefully factorized convolutions and aggressive regularization: 21.2% top-1 and 5.6% top-5 error for single frame evaluation using a network with a computational cost of 5 billion multiply-adds per inference and with using less than 25 million parameters.

****

Cross Modal Distillation for Supervision Transfer

Saurabh Gupta, Judy Hoffman, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2827-2836

In this work we propose a technique that transfers supervision between images from different modalities. We use learned representations from a large labeled modality as supervisory signal for training representations for a new unlabeled paired modality. Our method enables learning of rich representations for unlabeled modalities and can be used as a pre-training procedure for new modalities with limited labeled data. We transfer supervision from labeled RGB images to unlabeled depth and optical flow images and demonstrate large improvements for both these cross modal supervision transfers.

****

Efficient Point Process Inference for Large-Scale Object Detection

Trung T. Pham, Seyed Hamid Rezatofighi, Ian Reid, Tat-Jun Chin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2837-2845

We tackle the problem of large-scale object detection in images, where the number of objects can be arbitrarily large, and can exhibit significant overlap/occlusion. A successful approach to modelling the large-scale nature of this problem has been via point process density functions which jointly encode object qualities and spatial interactions. But the corresponding optimisation problem is typically difficult or intractable, and many of the best current methods rely on Monte Carlo Markov Chain (MCMC) simulation, which converges slowly in a large solution space. We propose an efficient point process inference for large-scale object detection using discrete energy minimization. In particular, we approximate the solution space by a finite set of object proposals and cast the point process density function to a corresponding energy function of binary variables whose values indicate which object proposals are accepted. We resort to the local submodular approximation (LSA) based trust-region optimisation to find the optimal solution. Furthermore we analyse the error of LSA approximation, and show how to adjust the point process energy to dramatically speed up the convergence without harms in the optimality. We demonstrate the superior efficiency and accuracy of our method using a variety of large-scale object detection applications such as crowd human detection, birds, cells counting/localization.

****

Weakly Supervised Deep Detection Networks
Hakan Bilen, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2016, pp. 2846-2854
Weakly supervised learning of object detection is an important problem in image understanding that still does not have a satisfactory solution. In this paper, we address this problem by exploiting the power of deep convolutional neural net works pre-trained on large-scale image-level classification tasks. We propose a weakly supervised deep detection architecture that modifies one such network to operate at the level of image regions, performing simultaneously region selectio n and classification. Trained as an image classifier, the architecture implicitl y learns object detectors that are better than alternative weakly supervised det ection systems on the PASCAL VOC data. The model, which is a simple and elegant end-to-end architecture, outperforms standard data augmentation and fine-tuning techniques for the task of image-level classification as well.
****

BORDER: An Oriented Rectangles Approach to Texture-Less Object Recognition
Jacob Chan, Jimmy Addison Lee, Qian Kemao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2855-2863
This paper presents an algorithm coined BORDER (Bounding Oriented-Rectangle Desc riptors for Enclosed Regions) for texture-less object recognition. By fusing a r egional object encompassment concept with descriptor-based pipelines, we extend local-patches into scalable object-sized oriented rectangles for optimal object information encapsulation with minimal outliers. We correspondingly introduce a modified line-segment detection technique termed Linelets to stabilize keypoint repeatability in homogenous conditions. In addition, a unique sampling technique facilitates the incorporation of robust angle primitives to produce discriminat ive rotation-invariant descriptors. BORDER's high competence in object recogniti on particularly excels in homogenous conditions obtaining superior detection rat es in the presence of high-clutter, occlusion and scale-rotation changes when co mpared with modern state-of-the-art texture-less object detectors such as BOLD a nd LINE2D on public texture-less object databases.
****

Active Image Segmentation Propagation
Suyog Dutt Jain, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2864-2873
We propose a semi-automatic method to obtain foreground object masks for a large set of related images. We develop a stagewise active approach to propagation: in each stage, we actively determine the images that appear most valuable for hu man annotation, then revise the foreground estimates in all unlabeled images acc ordingly. In order to identify images that, once annotated, will propagate well to other examples, we introduce an active selection procedure that operates on the joint segmentation graph over all images. It prioritizes human intervention for those images that are uncertain and influential in the graph, while also mu tually diverse. We apply our method to obtain foreground masks for over 1 milli on images. Our method yields state-of-the-art accuracy on the ImageNet and MIT Object Discovery datasets, and it focuses human attention more effectively than existing propagation strategies.
****

Inside-Outside Net: Detecting Objects in Context With Skip Pooling and Recurrent Neural Networks
Sean Bell, C. Lawrence Zitnick, Kavita Bala, Ross Girshick; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2874 -2883
It is well known that contextual and multi-scale representations are important f or accurate visual recognition. In this paper we present the Inside-Outside Net (ION), an object detector that exploits information both inside and outside the region of interest. Contextual information outside the region of interest is int egrated using spatial recurrent neural networks. Inside, we use skip pooling to extract information at multiple scales and levels of abstraction. Through extens ive experiments we evaluate the design space and provide readers with an overvie

w of what tricks of the trade are important. ION improves state-of-the-art on PA
SCAL VOC 2012 object detection from 73.9% to 77.9% mAP. On the new and more chal
lenging MS COCO dataset, we improve state-of-the-art from 19.7% to 33.1% mAP. In
 the 2015 MS COCO Detection Challenge, our ION model won "Best Student Entry" an
d finished 3rd place overall. As intuition suggests, our detection results provi
de strong evidence that context and multi-scale representations improve small ob
ject detection.
****

RIFD-CNN: Rotation-Invariant and Fisher Discriminative Convolutional Neural Netw
orks for Object Detection
Gong Cheng, Peicheng Zhou, Junwei Han; Proceedings of the IEEE Conference on Com
puter Vision and Pattern Recognition (CVPR), 2016, pp. 2884-2893
Thanks to the powerful feature representations obtained through deep convolution
al neural network (CNN), the performance of object detection has recently been s
ubstantially boosted. Despite the remarkable success, the problems of object rot
ation, within-class variability, and between-class similarity remain several maj
or challenges. To address these problems, this paper proposes a novel and effect
ive method to learn a rotation-invariant and Fisher discriminative CNN (RIFD-CNN
) model. This is achieved by introducing and learning a rotation-invariant layer
 and a Fisher discriminative layer, respectively, on the basis of the existing h
igh-capacity CNN architectures. Specifically, the rotation-invariant layer is tr
ained by imposing an explicit regularization constraint on the objective functio
n that enforces invariance on the CNN features before and after rotating. The Fi
sher discriminative layer is trained by imposing the Fisher discrimination crite
rion on the CNN features so that they have small within-class scatter but large
between-class separation. In the experiments, we comprehensively evaluate the pr
oposed method for object detection task on a public available aerial image datas
et and the PASCAL VOC 2007 dataset. State-of-the-art results are achieved compar
ed with the existing baseline methods.
****

Reinforcement Learning for Visual Object Detection
Stefan Mathe, Aleksis Pirinen, Cristian Sminchisescu; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2894-2902
One of the most widely used strategies for visual object detection is based on e
xhaustive spatial hypothesis search. While methods like sliding windows have bee
n successful and effective for many years, they are still brute-force, independe
nt of the image content and the visual category being searched. In this paper we
 present formally rigorous sequential models that accumulate evidence collected
at a small set of image locations in order to detect visual objects effectively.
 By formulating sequential search as reinforcement learning of the search policy
 (including the stopping condition), our fully trainable model can explicitly ba
lance for each class, specifically, the conflicting goals of exploration -- samp
ling more image regions for better accuracy --, and exploitation -- stopping the
 search efficiently when sufficiently confident in the target's location. The me
thodology is general and applicable to any detector response function. We report
 encouraging results in the PASCAL VOC 2012 object detection test set showing th
at the proposed methodology achieves almost two orders of magnitude speed-up ov
er sliding window methods.
****

Detecting Repeating Objects Using Patch Correlation Analysis
Inbar Huberman, Raanan Fattal; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2016, pp. 2903-2911
In this paper we describe a new method for detecting and counting a repeating ob
ject in an image. While the method relies on a fairly sophisticated deformable p
art model, unlike existing techniques it estimates the model parameters in an un
supervised fashion thus alleviating the need for a user-annotated training data
and avoiding the associated specificity. This automatic fitting process is carri
ed out by exploiting the recurrence of small image patches associated with the r
epeating object and analyzing their spatial correlation. The analysis allows us
to reject outlier patches, recover the visual and shape parameters of the part m

odel, and detect the object instances efficiently. In order to achieve a practi cal system which is able to cope with diverse images, we describe a simple and i ntuitive active-learning procedure that updates the object classification by que rying the user on very few carefully chosen marginal classifications. Evaluation of the new method against the state-of-the-art techniques demonstrates its abil ity to achieve higher accuracy through a better user experience.
****

Analyzing Classifiers: Fisher Vectors and Deep Neural Networks
Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, Wojciech Samek; Proceedings of the IEEE Conference on Computer Vision and Patter n Recognition (CVPR), 2016, pp. 2912-2920
Fisher vector (FV) classifiers and Deep Neural Networks (DNNs) are popular and s uccessful algorithms for solving image classification problems. However, both ar e generally considered `black box' predictors as the non-linear transformations involved have so far prevented transparent and interpretable reasoning. Recently , a principled technique, Layer-wise Relevance Propagation (LRP), has been devel oped in order to better comprehend the inherent structured reasoning of complex nonlinear classification models such as Bag of Feature models or DNNs. In this p aper we (1) extend the LRP framework also for Fisher vector classifiers and then use it as analysis tool to (2) quantify the importance of context for classific ation, (3) qualitatively compare DNNs against FV classifiers in terms of importa nt image regions and (4) detect potential flaws and biases in data. All experim ents are performed on the PASCAL VOC 2007 and ILSVRC 2012 data sets.
****

Learning Deep Features for Discriminative Localization
Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929
In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network (CNN) t o have remarkable localization ability despite being trained on image-level labe ls. While this technique was previously proposed as a means for regularizing tra ining, we find that it actually builds a generic localizable deep representation that exposes the implicit attention of CNNs on image. Despite the apparent simp licity of global average pooling, we are able to achieve 37.1% top-5 error for o bject localization on ILSVRC 2014 without training on any bounding box annotatio n. We demonstrate that our network is able to localize the discriminative image regions on a variety of tasks despite not being trained for them.
****

Seeing Through the Human Reporting Bias: Visual Classifiers From Noisy Human-Cen tric Labels
Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, Ross Girshick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2930-2939
When human annotators are given a choice about what to label in an image, they a pply their own subjective judgments on what to ignore and what to mention. We r efer to these noisy "human-centric" annotations as exhibiting human reporting bi as. Examples of such annotations include image tags and keywords found on photo sharing sites, or in datasets containing image captions. In this paper, we use these noisy annotations for learning visually correct image classifiers. Such an notations do not use consistent vocabulary, and miss a significant amount of the information present in an image; however, we demonstrate that the noise in thes e annotations exhibits structure and can be modeled. We propose an algorithm to decouple the human reporting bias from the correct visually grounded labels. Our results are highly interpretable for reporting "what's in the image" versus "wh at's worth saying." We demonstrate the algorithm's efficacy along a variety of metrics and datasets, including MS COCO and Yahoo Flickr 100M. We show significa nt improvements over traditional algorithms for both image classification and im age captioning, doubling the performance of existing methods in some cases.
****

Learning Aligned Cross-Modal Representations From Weakly Aligned Data
Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2940-2949
People can recognize scenes across many different modalities beyond natural images. In this paper, we investigate how to learn cross-modal scene representations that transfer across modalities. To study this problem, we introduce a new cross-modal scene dataset. While convolutional neural networks can categorize cross-modal scenes well, they also learn an intermediate representation not aligned across modalities, which is undesirable for cross-modal transfer applications. We present methods to regularize cross-modal convolutional neural networks so that they have a shared representation that is agnostic of the modality. Our experiments suggest that our scene representation can help transfer representations across modalities for retrieval. Moreover, our visualizations suggest that units emerge in the shared representation that tend to activate on consistent concepts independently of the modality.
****

A Probabilistic Collaborative Representation Based Approach for Pattern Classification
Sijia Cai, Lei Zhang, Wangmeng Zuo, Xiangchu Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2950-2959
Conventional representation based classifiers, ranging from the classical nearest neighbor classifier and nearest subspace classifier to the recently developed sparse representation based classifier (SRC) and collaborative representation based classifier (CRC), are essentially distance based classifiers. Though SRC and CRC have shown interesting classification results, their intrinsic classification mechanism remains unclear. In this paper we propose a probabilistic collaborative representation framework, where the probability that a test sample belongs to the collaborative subspace of all classes can be well defined and computed. Consequently, we present a probabilistic collaborative representation based classifier (ProCRC), which jointly maximizes the likelihood that a test sample belongs to each of the multiple classes. The final classification is performed by checking which class has the maximum likelihood. The proposed ProCRC has a clear probabilistic interpretation, and it shows superior performance to many popular classifiers, including SRC, CRC and SVM. Coupled with the CNN features, it also leads to state-of-the-art classification results on a variety of challenging visual datasets.
****

Learning Structured Inference Neural Networks With Label Relations
Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2960-2968
Images of scenes have various objects as well as abundant attributes, and diverse levels of visual categorization are possible. A natural image could be assigned with fine-grained labels that describe major components, coarse-grained labels that depict high level abstraction or a set of labels that reveal attributes. Such categorization at different concept layers can be modeled with label graphs encoding label information. In this paper, we exploit this rich information with a state-of-art deep learning framework, and propose a generic structured model that leverages diverse label relations to improve image classification performance. Our approach employs a novel stacked label prediction neural network, capturing both inter-level and intra-level label semantics. We evaluate our method on benchmark image datasets, and empirical results illustrate the efficacy of our model.
****

Discriminative Multi-Modal Feature Fusion for RGBD Indoor Scene Recognition
Hongyuan Zhu, Jean-Baptiste Weibel, Shijian Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2969-2976
RGBD scene recognition has attracted increasingly attention due to the rapid development of depth sensors and their wide application scenarios. While many resea

rch has been conducted, most work used hand-crafted features which are difficult to capture high-level semantic structures. Recently, the feature extracted from deep convolutional neural network has produced state-of-the-art results for various computer vision tasks, which inspire researchers to explore incorporating CNN learned features for RGBD scene understanding. On the other hand, most existing work combines rgb and depth features without adequately exploiting the consistency and complementary information between them. Inspired by some recent work on RGBD object recognition using multi-modal feature fusion, we introduce a novel discriminative multi-modal fusion framework for rgbd scene recognition for the first time which simultaneously considers the inter- and intra-modality correlation for all samples and meanwhile regularizing the learned features to be discriminative and compact. The results from the multimodal layer can be back-propagated to the lower CNN layers, hence the parameters of the CNN layers and multimodal layers are updated iteratively until convergence. Experiments on the recently proposed large scale SUN RGB-D datasets show that our method achieved the state-of-the-art without any image segmentation.
****

Conditional Graphical Lasso for Multi-Label Image Classification
Qiang Li, Maoying Qiao, Wei Bian, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2977-2986
Multi-label image classification aims to predict multiple labels for a single image which contains diverse content. By utilizing label correlations, various techniques have been developed to improve classification performance. However, current existing methods either neglect image features when exploiting label correlations or lack the ability to learn image-dependent conditional label structures. In this paper, we develop conditional graphical Lasso (CGL) to handle these challenges. CGL provides a unified Bayesian framework for structure and parameter learning conditioned on image features. We formulate the multi-label prediction as CGL inference problem, which is solved by a mean field variational approach. Meanwhile, CGL learning is efficient due to a tailored proximal gradient procedure by applying the maximum a posterior (MAP) methodology. CGL performs competitively for multi-label image classification on benchmark datasets MULAN scene, PASCAL VOC 2007 and PASCAL VOC 2012, compared with the state-of-the-art multi-label classification algorithms.
****

Region Ranking SVM for Image Classification
Zijun Wei, Minh Hoai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2987-2996
The success of an image classification algorithm largely depends on how it incorporates local information in the global decision. Popular approaches such as average-pooling and max-pooling are suboptimal in many situations. In this paper we propose Region Ranking SVM(RRSVM), a novel method for pooling local information from multiple regions. RRSVM exploits the correlation of local regions in an image, and it jointly learns a region evaluation function and a scheme for integrating multiple regions. Experiments on PASCAL VOC 2007, VOC 2012, and ILSVRC2014 datasets show that RRSVM outperforms the methods that use the same feature type and extract features from the same set of local regions. IRSVM achieves similar to or better than the state-of-the-art performance on all datasets.
****

Predicting Motivations of Actions by Leveraging Text
Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2997-3005
Understanding human actions is a key problem in computer vision. However, recognizing actions is only the first step of understanding what a person is doing. In this paper, we introduce the problem of predicting why a person has performed an action in images. This problem has many applications in human activity understanding, such as anticipating or explaining an action. To study this problem, we introduce a new dataset of people performing actions annotated with likely motivations. However, the information in an image alone may not be sufficient to auto

matically solve this task. Since humans can rely on their lifetime of experiences to infer motivation, we propose to give computer vision systems access to some of these experiences by using recently developed natural language models to mine knowledge stored in massive amounts of text. While we are still far away from fully understanding motivation, our results suggest that transferring knowledge from language into vision can help machines understand why people in images might be performing an action.
****

BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition
Jakub Sochor, Adam Herout, Jiri Havel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3006-3015
We are dealing with the problem of fine-grained vehicle make&model recognition and verification. Our contribution is showing that extracting additional data from the video stream - besides the vehicle image itself - and feeding it into the deep convolutional neural network boosts the recognition performance considerably. This additional information includes: 3D vehicle bounding box used for "unpacking" the vehicle image, its rasterized low-resolution shape, and information about the 3D vehicle orientation. Experiments show that adding such information decreases classification error by 26% (the accuracy is improved from 0.772 to 0.832) and boosts verification average precision by 208% (0.378 to 0.785) compared to baseline pure CNN without any input modifications. Also, the pure baseline CNN outperforms the recent state of the art solution by 0.081. We provide an annotated set "BoxCars" of surveillance vehicle images augmented by various automatically extracted auxiliary information. Our approach and the dataset can considerably improve the performance of traffic surveillance systems.
****

Highway Vehicle Counting in Compressed Domain
Xu Liu, Zilei Wang, Jiashi Feng, Hongsheng Xi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3016-3024
This paper presents a highway vehicle counting method in compressed domain, aiming at achieving acceptable estimation performance approaching the pixel-domain methods. Such a task essentially is challenging because the available information (e.g. motion vector) to describe vehicles in videos is quite limited and inaccurate, and the vehicle count in realistic traffic scenes always varies greatly. To tackle this issue, we first develop a batch of low-level features, which can be extracted from the encoding metadata of videos, to mitigate the informational insufficiency of compressed videos. Then we propose a Hierarchical Classification based Regression (HCR) model to estimate the vehicle count from features. HCR hierarchically divides the traffic scenes into different cases according to vehicle density, such that the broad-variation characteristics of traffic scenes can be better approximated. Finally, we evaluated the proposed method on the real highway surveillance videos. The results show that our method is very competitive to the pixel-domain methods, which can reach similar performance along with its lower complexity.
****

Camera Calibration From Periodic Motion of a Pedestrian
Shiyao Huang, Xianghua Ying, Jiangpeng Rong, Zeyu Shang, Hongbin Zha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3025-3033
Camera calibration directly from image sequences of a pedestrian without using any calibration object is a really challenging task and should be well solved in computer vision, especially in visual surveillance. In this paper, we propose a novel camera calibration method based on recovering the three orthogonal vanishing points (TOVPs), just using an image sequence of a pedestrian walking in a straight line, without any assumption of scenes or motions, e.g., control points with known 3D coordinates, parallel or perpendicular lines, non-natural or pre-designed special human motions, as often necessary in previous methods. The traces of shoes of a pedestrian carry more rich and easily detectable metric information than all other body parts in the periodic motion of a pedestrian, but such information is usually overlooked by previous work. In this paper, we employ the im

ages of the toes of the shoes on the ground plane to determine the vanishing poi nt corresponding to the walking direction, and then utilize harmonic conjugate p roperties in projective geometry to recover the vanishing point corresponding to the perpendicular direction of the walking direction in the horizontal plane an d the vanishing point corresponding to the vertical direction. After recovering all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be de termined. Experiments on various scenes and viewing angles prove the feasibility and accuracy of the proposed method.
****

Dynamic Image Networks for Action Recognition
Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, Stephen Gould; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C VPR), 2016, pp. 3034-3042
We introduce the concept of dynamic image, a novel compact representation of vid eos useful for video analysis especially when convolutional neural networks (CNN s) are used. The dynamic image is based on the rank pooling concept and is obtai ned through the parameters of a ranking machine that encodes the temporal evolut ion of the frames of the video. Dynamic images are obtained by directly applying rank pooling on the raw image pixels of a video producing a single RGB image pe r video. This idea is simple but powerful as it enables the use of existing CNN models directly on video data with fine-tuning. We present an efficient and effe ctive approximate rank pooling operator, speeding it up orders of magnitude comp ared to rank pooling. Our new approximate rank pooling CNN layer allows us to ge neralize dynamic images to dynamic feature maps and we demonstrate the power of our new representations on standard benchmarks in action recognition achieving s tate-of-the-art performance.
****

Detecting Events and Key Actors in Multi-Person Videos
Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin M urphy, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pat tern Recognition (CVPR), 2016, pp. 3043-3053
Multi-person event recognition is a challenging task, often with many people act ive in the scene but only a small subset contributing to an actual event. In thi s paper, we propose a model which learns to detect events in such videos while a utomatically "attending" to the people responsible for the event. Our model doe s not use explicit annotations regarding who or where those people are during tr aining and testing. In particular, we track people in videos and use a recurrent neural network (RNN) to represent the track features. We learn time-varying at tention weights to combine these features at each time-instant. The attended fea tures are then processed using another RNN for event detection/classification. Since most video datasets with multiple people are restricted to a small number of videos, we also collected a new basketball dataset comprising 257 basketball games with 14K event annotations corresponding to 11 event classes. Our model o utperforms state-of-the-art methods for both event classification and detectio n on this new dataset. Additionally, we show that the attention mechanism is abl e to consistently localize the relevant players.
****

Regularizing Long Short Term Memory With 3D Human-Skeleton Sequences for Action Recognition
Behrooz Mahasseni, Sinisa Todorovic; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2016, pp. 3054-3062
This paper argues that large-scale action recognition in video can be greatly im proved by providing an additional modality in training data -- namely, 3D human-skeleton sequences -- aimed at complementing poorly represented or missing featu res of human actions in the training videos. For recognition, we use Long Short Term Memory (LSTM) grounded via a deep Convolutional Neural Network (CNN) onto t he video. Training of LSTM is regularized using the output of another encoder LS TM (eLSTM) grounded on 3D human-skeleton training data. For such regularized tra ining of LSTM, we modify the standard backpropagation through time (BPTT) in ord er to address the well-known issues with gradient descent in constraint optimiza

tion. Our evaluation on three benchmark datasets -- Sports-1M, HMDB-51, and UCF101 -- shows accuracy improvements from 5.3% up to 17.4% relative to the state of the art.
****

Personalizing Human Video Pose Estimation
James Charles, Tomas Pfister, Derek Magee, David Hogg, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3063-3072
We propose a personalized ConvNet pose estimator that automatically adapts itself to the uniqueness of a person's appearance to improve pose estimation in long videos. We make the following contributions: (i) we show that given a few high-precision pose annotations, e.g. from a generic ConvNet pose estimator, additional annotations can be generated throughout the video using a combination of image-based matching for temporally distant frames, and dense optical flow for temporally local frames; (ii) we develop an occlusion aware self-evaluation model that is able to automatically select the high-quality and reject the erroneous additional annotations; and (iii) we demonstrate that these high-quality annotations can be used to fine-tune a ConvNet pose estimator and thereby personalize it to lock on to key discriminative features of the person's appearance. The outcome is a substantial improvement in the pose estimates for the target video using the personalized ConvNet compared to the original generic ConvNet. Our method outperforms the state of the art (including top ConvNet methods) by a large margin on three standard benchmarks, as well as on a new challenging YouTube video dataset. Furthermore, we show that training from the automatically generated annotations can be used to improve the performance of a generic ConvNet on other benchmarks.
****

End-To-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation
Wei Yang, Wanli Ouyang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3073-3082
Recently, Deep Convolutional Neural Networks (DCNNs) have been applied to the task of human pose estimation, and have shown its potential of learning better feature representations and capturing contextual relationships. However, it is difficult to incorporate domain prior knowledge such as geometric relationships among body parts into DCNNs. In addition, training DCNN-based body part detectors without consideration of global body joint consistency introduces ambiguities, which increases the complexity of training. In this paper, we propose a novel end-to-end framework for human pose estimation that combines DCNNs with the expressive deformable mixture of parts. We explicitly incorporate domain prior knowledge into the framework, which greatly regularizes the learning process and enables the flexibility of our framework for loopy models or tree-structured models. The effectiveness of jointly learning a DCNN with a deformable mixture of parts model is evaluated through intensive experiments on several widely used benchmarks. The proposed approach significantly improves the performance compared with state-of-the-art approaches, especially on benchmarks with challenging articulations.
****

Actor-Action Semantic Segmentation With Grouping Process Models
Chenliang Xu, Jason J. Corso; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3083-3092
Actor-action semantic segmentation made an important step toward advanced video understanding: what action is happening; who is performing the action; and where is the action happening in space-time. Current methods based on layered CRFs for this problem are local and unable to capture the long-ranging interactions of video parts. We propose a new model that combines the labeling CRF with a supervoxel hierarchy, where supervoxels at various scales provide cues for possible groupings of nodes in the CRF to encourage adaptive and long-ranging interactions. The new model defines a dynamic and continuous process of information exchange: the CRF influences what supervoxels in the hierarchy are active, and these acti

ve supervoxels, in turn, affect the connectivities in the CRF; we hence call it a grouping process model. By further incorporating the video-level recognition, the proposed method achieves a large margin of 60% relative improvement over the state of the art on the recent A2D large-scale video labeling dataset, which demonstrates the effectiveness of our modeling.

****

Temporal Action Localization With Pyramid of Score Distribution Features
Jun Yuan, Bingbing Ni, Xiaokang Yang, Ashraf A. Kassim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3093-3102

We investigate the feature design and classification architectures in temporal action localization. This application focuses on detecting and labeling actions in untrimmed videos, which brings more challenge than classifying pre-segmented videos. The major difficulty for action localization is the uncertainty of action occurrence and utilization of information from different scales. Two innovations are proposed to address this issue. First, we propose a Pyramid of Score Distribution Feature (PSDF) to capture the motion information at multiple resolutions centered at each detection window. This novel feature mitigates the influence of unknown action position and duration, and shows significant performance gain over previous detection approaches. Second, inter-frame consistency is further explored by incorporating PSDF into the state-of-the-art Recurrent Neural Networks, which gives additional performance gain in detecting actions in temporally untrimmed videos. We tested our action localization framework on the THUMOS'15 and MPII Cooking Activities Dataset, both of which show a large performance improvement over previous attempts.

****

Recognizing Activities of Daily Living With a Wrist-Mounted Camera
Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3103-3111

We present a novel dataset and a novel algorithm for recognizing activities of daily living (ADL) from a first-person wearable camera. Handled objects are crucially important for egocentric ADL recognition. For specific examination of objects related to users' actions separately from other objects in an environment, many previous works have addressed the detection of handled objects in images captured from head-mounted and chest-mounted cameras. Nevertheless, detecting handled objects is not always easy because they tend to appear small in images. They can be occluded by a user's body. As described herein, we mount a camera on a user's wrist. A wrist-mounted camera can capture handled objects at a large scale, and thus it enables us to skip the object detection process. To compare a wrist-mounted camera and a head-mounted camera, we also developed a novel and publicly available dataset that includes videos and annotations of daily activities captured simultaneously by both cameras. Additionally, we propose a discriminative video representation that retains spatial and temporal information after encoding the frame descriptors extracted by convolutional neural networks.

****

Harnessing Object and Scene Semantics for Large-Scale Video Understanding
Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3112-3121

Large-scale action recognition and video categorization are important problems in computer vision. To address these problems, we propose a novel object- and scene-based semantic fusion network and representation. Our semantic fusion network combines three streams of information using a three-layer neural network: (i) frame-based low-level CNN features, (ii) object features from a state-of-the-art large-scale CNN object-detector trained to recognize 20K classes, and (iii) scene features from a state-of-the-art CNN scene-detector trained to recognize 205 scenes. The trained network achieves improvements in supervised activity and video categorization in two complex large-scale datasets - ActivityNet and FCVID, respectively. Further, by examining and back propagating information through the fusion network, semantic relationships (correlations) between video classes and

objects/scenes can be discovered. These video class-object/video class-scene relationships can in turn be used as semantic representation for the video classes themselves. We illustrate effectiveness of this semantic representation through experiments on zero-shot action/video classification and clustering.
****

Video-Story Composition via Plot Analysis
Jinsoo Choi, Tae-Hyun Oh, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3122-3130
We address the problem of composing a story out of multiple short video clips taken by a person during an activity or experience. Inspired by plot analysis of written stories, our method generates a sequence of video clips ordered in such a way that it reflects plot dynamics and content coherency. That is, given a set of multiple video clips, our method composes a video which we call a video-story. We define metrics on scene dynamics and coherency by dense optical flow features and a patch matching algorithm. Using these metrics, we define an objective function for the video-story. To efficiently search for the best video-story, we introduce a novel Branch-and-Bound algorithm which guarantees the global optimum. We collect the dataset consisting of 23 video sets from the web, resulting in a total of 236 individual video clips. With the acquired dataset, we perform extensive user studies involving 30 human subjects by which the effectiveness of our approach is quantitatively and qualitatively verified.
****

Temporal Action Detection Using a Statistical Language Model
Alexander Richard, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3131-3140
While current approaches to action recognition on pre-segmented video clips already achieve high accuracies, temporal action detection is still far from comparably good results. Automatically locating and classifying the relevant action segments in videos of varying lengths proves to be a challenging task. We propose a novel method for temporal action detection including statistical length and language modeling to represent temporal and contextual structure. Our approach aims at globally optimizing the joint probability of three components, a length and language model and a discriminative action model, without making intermediate decisions. The problem of finding the most likely action sequence and the corresponding segment boundaries in an exponentially large search space is addressed by dynamic programming. We provide an extensive evaluation of each model component on Thumos 14, a large action detection dataset, and report state-of-the-art results on three datasets.
****

Multi-Scale Patch Aggregation (MPA) for Simultaneous Detection and Segmentation
Shu Liu, Xiaojuan Qi, Jianping Shi, Hong Zhang, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3141-3149
Aiming at simultaneous detection and segmentation (SDS), we propose a proposal-free framework, which detect and segment object instances via mid-level patches. We design a unified trainable network on patches, which is followed by a fast and effective patch aggregation algorithm to infer object instances. Our method benefits from end-to-end training. Without object proposal generation, computation time can also be reduced. In experiments, our method yields results 62.1% and 61.8% in terms of mAPr on VOC2012 segmentation val and VOC2012 SDS val, which are state-of-the-art at the time of submission. We also report results on Microsoft COCO test-std/test-dev dataset in this paper.
****

Instance-Aware Semantic Segmentation via Multi-Task Network Cascades
Jifeng Dai, Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3150-3158
Semantic segmentation research has recently witnessed rapid progress, but many leading methods are unable to identify object instances. In this paper, we present Multi-task Network Cascades for instance-aware semantic segmentation. Our model consists of three networks, respectively differentiating instances, estimating

masks, and categorizing objects. These networks form a cascaded structure, and are designed to share their convolutional features. We develop an algorithm for the nontrivial end-to-end training of this causal, cascaded structure. Our solution is a clean, single-step training framework and can be generalized to cascades that have more stages. We demonstrate state-of-the-art instance-aware semantic segmentation accuracy on PASCAL VOC. Meanwhile, our method takes only 360ms testing an image using VGG-16, which is two orders of magnitude faster than previous systems for this challenging problem. As a by product, our method also achieves compelling object detection results which surpass the competitive Fast/Faster R-CNN systems. The method described in this paper is the foundation of our submissions to the MS COCO 2015 segmentation competition, where we won the 1st place.

****

ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation

Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3159-3167

Large-scale data are of crucial importance for learning semantic segmentation models, but annotating per-pixel masks is a tedious and inefficient procedure. We note that for the topic of interactive image segmentation, scribbles are very widely used in academic research and commercial software, and are recognized as one of the most user-friendly ways of interacting. In this paper, we propose to use scribbles to annotate images, and develop an algorithm to train convolutional networks for semantic segmentation supervised by scribbles. Our algorithm is based on a graphical model that jointly propagates information from scribbles to unmarked pixels and learns network parameters. We present competitive object semantic segmentation results on the PASCAL VOC dataset by using scribbles as annotations. Scribbles are also favored for annotating stuff (e.g., water, sky, grass) that has no well-defined shape, and our method shows excellent results on the PASCAL-CONTEXT dataset thanks to extra inexpensive scribble annotations.

****

Feature Space Optimization for Semantic Video Segmentation

Abhijit Kundu, Vibhav Vineet, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3168-3175

We present an approach to long-range spatio-temporal regularization in semantic video segmentation. Temporal regularization in video is challenging because both the camera and the scene may be in motion. Thus Euclidean distance in the space-time volume is not a good proxy for correspondence. We optimize the mapping of pixels to a Euclidean feature space so as to minimize distances between corresponding points. Structured prediction is performed by a dense CRF that operates on the optimized features. Experimental results demonstrate that the presented approach increases the accuracy and temporal consistency of semantic video segmentation.

****

Large-Scale Semantic 3D Reconstruction: An Adaptive Multi-Resolution Model for Multi-Class Volumetric Labeling

Maros Blaha, Christoph Vogel, Audrey Richard, Jan D. Wegner, Thomas Pock, Konrad Schindler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3176-3184

We propose an adaptive multi-resolution formulation of semantic 3D reconstruction. Given a set of images of a scene, semantic 3D reconstruction aims to densely reconstruct both the 3D shape of the scene and a segmentation into semantic object classes. Jointly reasoning about shape and class allows one to take into account class-specific shape priors (e.g., building walls should be smooth and vertical, and vice versa smooth, vertical surfaces are likely to be building walls), leading to improved reconstruction results. So far, semantic 3D reconstruction methods have been limited to small scenes and low resolution, because of their large memory footprint and computational cost. To scale them up to large scenes, we propose a hierarchical scheme which refines the reconstruction only in regions that are likely to contain a surface, exploiting the fact that both high spatia

l resolution and high numerical precision are only required in those regions. Ou
r scheme amounts to solving a sequence of convex optimizations while progressive
ly removing constraints, in such a way that the energy, in each iteration, is th
e tightest possible approximation of the underlying energy at full resolution. I
n our experiments the method saves up to 98% memory and 95% computation time, wi
thout any loss of accuracy.
****

Semantic Object Parsing With Local-Global Long Short-Term Memory
Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, Shuicheng Ya
n; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
 (CVPR), 2016, pp. 3185-3193
Semantic object parsing is a fundamental task for understanding objects in detai
l in computer vision community, where incorporating multi-level contextual infor
mation is critical for achieving such fine-grained pixel-level recognition. Prio
r methods often leverage the contextual information through post-processing pred
icted confidence maps. In this work, we propose a novel deep Local-Global Long S
hort-Term Memory (LG-LSTM) architecture to seamlessly incorporate short-distance
 and long-distance spatial dependencies into the feature learning over all pixel
 positions. In each LG-LSTM layer, local guidance from neighboring positions and
 global guidance from the whole image are imposed on each position to better exp
loit complex local and global contextual information. Individual LSTMs for disti
nct spatial dimensions are also utilized to intrinsically capture various spatia
l layouts of semantic parts in the images, yielding distinct hidden and memory c
ells of each position for each dimension. In our parsing approach, several LG-LS
TM layers are stacked and appended to the intermediate convolutional layers to d
irectly enhance visual features, allowing network parameters to be learned in an
 end-to-end way. The long chains of sequential computation by stacked LG-LSTM la
yers also enable each pixel to sense a much larger region for inference benefiti
ng from the memorization of previous dependencies in all positions along all dim
ensions. Comprehensive evaluations on three public datasets well demonstrate the
 significant superiority of our LG-LSTM over other state-of-the-art methods for
object parsing.
****

Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation
Guosheng Lin, Chunhua Shen, Anton van den Hengel, Ian Reid; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3194
-3203
Recent advances in semantic image segmentation have mostly been achieved by trai
ning deep convolutional neural networks(CNNs). We show how to improve semantic s
egmentation through the use of contextual information; specifically, we explore
'patch-patch' context between image regions, and 'patch-background' context. For
 learning from the patch-patch context, we formulate Conditional Random Fields (
CRFs) with CNN-based pairwise potential functions to capture semantic correlatio
ns between neighboring patches. Efficient piecewise training of the proposed dee
p structured model is then applied to avoid repeated expensive CRF inference for
 back propagation. For capturing the patch-background context, we show that a ne
twork design with traditional multi-scale image input and sliding pyramid poolin
g is effective for improving performance. Our experimental results set new state
-of-the-art performance on a number of popular semantic segmentation datasets, i
ncluding NYUDv2, PASCAL VOC 2012, PASCAL-Context, and SIFT-flow. In particular,
we achieve an intersection-over-union score of 78.0 on the challenging PASCAL VO
C 2012 dataset.
****

Learning Transferrable Knowledge for Semantic Segmentation With Deep Convolution
al Neural Network
Seunghoon Hong, Junhyuk Oh, Honglak Lee, Bohyung Han; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3204-3212
We propose a novel weakly-supervised semantic segmentation algorithm based on De
ep Convolutional Neural Net- work (DCNN). Contrary to existing weakly-supervised
 approaches, our algorithm exploits auxiliary segmentation an- notations availab

le for different categories to guide segmentations on images with only image-lev
el class labels. To make segmentation knowledge transferrable across categories,
 we design a decoupled encoder-decoder architecture with attention model. In thi
s architecture, the model generates spatial highlights of each category presente
d in images using an attention model, and subsequently per- forms binary segment
ation for each highlighted region using decoder. Combining attention model, the
decoder trained with segmentation annotations in different categories boosts acc
uracy of weakly-supervised semantic segmentation. The proposed algorithm demonst
rates substantially improved performance compared to the state-of-the- art weakl
y-supervised techniques in PASCAL VOC 2012 dataset when our model is trained wit
h the annotations in 60 exclusive categories in Microsoft COCO dataset.
****

The Cityscapes Dataset for Semantic Urban Scene Understanding
Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, R
odrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele; Proceedings of the IEEE
 Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213-32
23
Visual understanding of complex urban street scenes is an enabling factor for a
wide range of applications. Object detection has benefited enormously from large
-scale datasets, especially in the context of deep learning. For semantic urban
scene understanding, however, no current dataset adequately captures the complex
ity of real-world urban scenes. To address this, we introduce Cityscapes, a ben
chmark suite and large-scale dataset to train and test approaches for pixel-leve
l and instance-level semantic labeling. Cityscapes is comprised of a large, dive
rse set of stereo video sequences recorded in streets from 50 different cities.
5000 of these images have high quality pixel-level annotations; 20000 additional
 images have coarse annotations to enable methods that leverage large volumes of
 weakly-labeled data. Crucially, our effort exceeds previous attempts in terms o
f dataset size, annotation richness, scene variability, and complexity. Our acco
mpanying empirical study provides an in-depth analysis of the dataset characteri
stics, as well as a performance evaluation of several state-of-the-art approache
s based on our benchmark.
****

Gaussian Conditional Random Field Network for Semantic Segmentation
Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, Rama Chellapa; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3
224-3233
In contrast to the existing approaches that use discrete Conditional Random Fiel
d (CRF) models, we propose to use a Gaussian CRF model for the task of semantic
segmentation. We propose a novel deep network, which we refer to as Gaussian Mea
n Field (GMF) network, whose layers perform mean field inference over a Gaussian
 CRF. The proposed GMF network has the desired property that each of its layers
produces an output that is closer to the maximum a posteriori solution of the Ga
ussian CRF compared to its input. By combining the proposed GMF network with dee
p Convolutional Neural Networks (CNNs), we propose a new end-to-end trainable Ga
ussian conditional random field network. The proposed Gaussian CRF network is co
mposed of three sub-networks: (i) a CNN-based unary network for generating unary
 potentials, (ii) a CNN-based pairwise network for generating pairwise potential
s, and (iii) a GMF network for performing Gaussian CRF inference. When trained e
nd-to-end in a discriminative fashion, and evaluated on the challenging PASCALVO
C 2012 segmentation dataset, the proposed Gaussian CRF network outperforms vario
us recent semantic segmentation approaches that combine CNNs with discrete CRF m
odels.
****

The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segment
ation of Urban Scenes
German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, Antonio M. Lopez;
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C
VPR), 2016, pp. 3234-3243
Vision-based semantic segmentation in urban scenarios is a key functionality for

autonomous driving. Recent revolutionary results of deep convolutional neural networks (DCNNs) foreshadow the advent of reliable classifiers to perform such visual tasks. However, DCNNs require learning of many parameters from raw images; thus, having a sufficient amount of diverse images with class annotations is needed. These annotations are obtained via cumbersome, human labour which is particularly challenging for semantic segmentation since pixel-level annotations are required. In this paper, we propose to use a virtual world to automatically generate realistic synthetic images with pixel-level annotations. Then, we address the question of how useful such data can be for semantic segmentation -- in particular, when using a DCNN paradigm. In order to answer this question we have generated a synthetic collection of diverse urban images, named SYNTHIA, with automatically generated class annotations. We use SYNTHIA in combination with publicly available real-world urban images with manually provided annotations. Then, we conduct experiments with DCNNs that show how the inclusion of SYNTHIA in the training stage significantly improves performance on the semantic segmentation task.

****

Progressive Prioritized Multi-View Stereo

Alex Locher, Michal Perdoch, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3244-3252

This work proposes a progressive patch based multi-view stereo algorithm able to deliver a dense point cloud at any time. This enables an immediate feedback on the reconstruction process in a user centric scenario. With increasing processing time, the model is improved in terms of resolution and accuracy. The algorithm explicitly handles input images with varying effective scale and creates visually pleasing point clouds. A priority scheme assures that the limited computational power is invested in scene parts, where the user is most interested in or the overall error can be reduced the most. The architecture of the proposed pipeline allows fast processing times in large scenes using a pure open-source CPU implementation. We show the performance of our algorithm on challenging standard datasets as well as on real-world scenes and compare it to the baseline.

****

WarpNet: Weakly Supervised Matching for Single-View Reconstruction

Angjoo Kanazawa, David W. Jacobs, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3253-3261

We present an approach to matching images of objects in fine-grained datasets without using part annotations, with an application to the challenging problem of weakly supervised single-view reconstruction. This is in contrast to prior works that require part annotations, since matching objects across class and pose variations is challenging with appearance features alone. We overcome this challenge through a novel deep learning architecture, WarpNet, that aligns an object in one image with a different object in another. We exploit the structure of the fine-grained dataset to create artificial data for training this network in an unsupervised-discriminative learning approach. The output of the network acts as a spatial prior that allows generalization at test time to match real images across variations in appearance, viewpoint and articulation. On the CUB-200-2011 dataset of bird categories, we improve the AP over an appearance-only network by 13.6%. We further demonstrate that our WarpNet matches, together with the structure of fine-grained datasets, allow single-view reconstructions with quality comparable to using annotated point correspondences.

****

What Sparse Light Field Coding Reveals About Scene Structure

Ole Johannsen, Antonin Sulc, Bastian Goldluecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3262-3270

In this paper, we propose a novel method for depth estimation in light fields which employs a specifically designed sparse decomposition to leverage the depth-orientation relationship on its epipolar plane images. The proposed method learns the structure of the central view and uses this information to construct a light field dictionary for which groups of atoms correspond to unique disparities. This dictionary is then used to code a sparse representation of the light field.

Analysing the coefficients of this representation with respect to the disparities of their corresponding atoms yields an accurate and robust estimate of depth. In addition, if the light field has multiple depth layers, such as for reflective or transparent surfaces, statistical analysis of the coefficients can be employed to infer the respective depth of the superimposed layers.
****

Online Reconstruction of Indoor Scenes From RGB-D Streams
Hao Wang, Jun Wang, Wang Liang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3271-3279
A system capable of performing robust online volumetric reconstruction of indoor scenes based on input from a handheld RGB-D camera is presented. Our system is powered by a two-pass reconstruction scheme. The first pass tracks camera poses at video rate and simultaneously constructs a pose graph on-the-fly. The tracker operates in real-time, which allows the reconstruction results to be visualized during the scanning process. Live visual feedbacks makes the scanning operation fast and intuitive. Upon termination of scanning, the second pass takes place to handle loop closures and reconstruct the final model using globally refined camera trajectories. The system is online with low delay and returns a dense model of sufficient accuracy. The beauty of this system lies in its speed, accuracy, simplicity and ease of implementation when compared to previous methods. We demonstrate the performance of our system on several real-world scenes and quantitatively assess the modeling accuracy with respect to ground truth models obtained from a LIDAR scanner.
****

Patches, Planes and Probabilities: A Non-Local Prior for Volumetric 3D Reconstruction
Ali Osman Ulusoy, Michael J. Black, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3280-3289
In this paper, we propose a non-local structured prior for volumetric multi-view 3D reconstruction. Towards this goal, we present a novel Markov random field model based on ray potentials in which assumptions about large 3D surface patches such as planarity or Manhattan world constraints can be efficiently encoded as probabilistic priors. We further derive an inference algorithm that reasons jointly about voxels, pixels and image segments, and estimates marginal distributions of appearance, occupancy, depth, normals and planarity. Key to tractable inference is a novel hybrid representation that spans both voxel and pixel space and that integrates non-local information from 2D image segmentations in a principled way. We compare our non-local prior to commonly employed local smoothness assumptions and a variety of state-of-the-art volumetric reconstruction baselines on challenging outdoor scenes with textureless and reflective surfaces. Our experiments indicate that regularizing over larger distances has the potential to resolve ambiguities where local regularizers fail.
****

Single Image Camera Calibration With Lenticular Arrays for Augmented Reality
Ian Schillebeeckx, Robert Pless; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3290-3298
We consider the problem of camera pose estimation for a scenario where the camera may have continuous and unknown changes in its focal length. Understanding frame by frame changes in camera focal length is vital to accurately estimating camera pose and vital to accurately render virtual objects in a scene with the correct perspective. However, most approaches to camera calibration require geometric constraints from many frames or the observation of a 3D calibration object --- both of which may not be feasible in augmented reality settings. This paper introduces a calibration objects based on a flat lenticular array that creates a color coded light-field whose observed color changes depending on the angle from which it is viewed. We derive an approach to estimate the focal length of the camera and the relative pose of an object from a single image. We characterize the performance of camera calibration across various focal lengths and camera models, and we demonstrate the advantages of the focal length estimation in rendering a virtual object in a video with constant zooming.

****
Augmented Blendshapes for Real-Time Simultaneous 3D Head Modeling and Facial Mot
ion Capture
Diego Thomas, Rin-ichiro Taniguchi; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2016, pp. 3299-3308
We propose a method to build in real-time animated 3D head models using a consum
er-grade RGB-D camera. Our framework is the first one to provide simultaneously
comprehensive facial motion tracking and a detailed 3D model of the user's head.
 Anyone's head can be instantly reconstructed and his facial motion captured wit
hout requiring any training or pre-scanning. The user starts facing the camera w
ith a neutral expression in the first frame, but is free to move, talk and chang
e his face expression as he wills otherwise. The facial motion is tracked using
a blendshape representation while the fine geometric details are captured using
a Bump image mapped over the template mesh. We propose an efficient algorithm to
 grow and refine the 3D model of the head on-the-fly and in real-time. We demons
trate robust and high-fidelity simultaneous facial motion tracking and 3D head m
odeling results on a wide range of subjects with various head poses and facial e
xpressions. Our proposed method offers interesting possibilities for animation p
roduction and 3D video telecommunications.
****
Learned Binary Spectral Shape Descriptor for 3D Shape Correspondence
Jin Xie, Meng Wang, Yi Fang; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2016, pp. 3309-3317
Dense 3D shape correspondence is an important problem in computer vision and com
puter graphics. Recently, the local shape descriptor based 3D shape corresponden
ce approaches have been widely studied, where the local shape descriptor is a re
al-valued vector to characterize the geometrical structure of the shape. Differe
nt from these real-valued local shape descriptors, in this paper, we propose to
learn a novel binary spectral shape descriptor with the deep neural network for
3D shape correspondence. The binary spectral shape descriptor can require less s
torage space and enable fast matching. First, based on the eigenvectors of the L
aplace-Beltrami operator, we construct a neural network to form a nonlinear spec
tral representation to characterize the shape. Then, for the defined positive an
d negative points on the shapes, we train the constructed neural network by mini
mizing the errors between the outputs and their corresponding binary descriptors
, minimizing the variations of the outputs of the positive points and maximizing
 the variations of the outputs of the negative points, simultaneously. Finally,
we binarize the output of the neural network to form the binary spectral shape d
escriptor for shape correspondence. The proposed binary spectral shape descripto
r is evaluated on the SCAPE and TOSCA 3D shape datasets for shape correspondenc
e. The experimental results demonstrate the effectiveness of the proposed binary
 shape descriptor for the shape correspondence task.
****
Multiple Model Fitting as a Set Coverage Problem
Luca Magri, Andrea Fusiello; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2016, pp. 3318-3326
This paper deals with the extraction of multiple models from noisy or outlier-co
ntaminated data. We cast the multi-model fitting problem in terms of set coveri
ng, deriving a simple and effective method that generalizes Ransac to multiple
models and deals with intersecting structures and outliers in a straightforwar
d and principled manner, while avoiding the typical shortcomings of sequential a
pproaches and those of clustering. The method compares favourably against the st
ate-of-the-art on simulated and publicly available real datasets.
****
Piecewise-Planar 3D Approximation From Wide-Baseline Stereo
Cedric Verleysen, Christophe De Vleeschouwer; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3327-3336
This paper approximates the 3D geometry of a scene by a small number of 3D plane
s. The method is especially suited to man-made scenes, and only requires two cal
ibrated wide-baseline views as inputs. It relies on the computation of a dense b

ut noisy 3D point cloud, as for example obtained by matching DAISY descriptors b etween the views. It then segments one of the two reference images, and adopts a multi-model fitting process to assign a 3D plane to each region, when the regi on is not detected as occluded. A pool of 3D plane hypotheses is first derived f rom the 3D point cloud, to include planes that reasonably approximate the part o f the 3D point cloud observed from each reference view between randomly selected triplets of 3D points. The hypothesis-to-region assignment problem is then form ulated as an energy-minimization problem, which simultaneously optimizes an orig inal data-fidelity term, the assignment smoothness over neighboring regions, and the number of assigned planar proxies. The synthesis of intermediate viewpoints demonstrates the effectiveness of our 3D reconstruction, and thereby the releva nce of our proposed data-fidelity metric.
****

Sparse to Dense 3D Reconstruction From Rolling Shutter Images
Olivier Saurer, Marc Pollefeys, Gim Hee Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3337-3345
It is well known that the rolling shutter effect in images captured with a movin g rolling shutter camera causes inaccuracies to 3D reconstructions. The problem is further aggravated with weak visual connectivity from wide baseline images ca ptured with a fast moving camera. In this paper, we propose and implement a pipe line for sparse to dense 3D construction with wide baseline images captured from a fast moving rolling shutter camera. pecifically, we propose a cost function f or Bundle Adjustment (BA) that models the rolling shutter effect, incorporates G PS/INS readings, and enforces pairwise smoothness between neighboring poses. We optimize over the 3D structures, camera poses and velocities. We also introduce a novel interpolation scheme for the rolling shutter plane sweep stereo algorith m that allows us to achieve a 7x speed up in the depth map computations for dens e reconstruction without losing accuracy. We evaluate our proposed pipeline over a 2.6km image sequence captured with a rolling shutter camera mounted on a movi ng car.
****

Consistency of Silhouettes and Their Duals
Matthew Trager, Martial Hebert, Jean Ponce; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3346-3354
Silhouettes provide rich information on three-dimensional shape, since the inter section of the associated visual cones generates the "visual hull", which enclos es and approximates the original shape. However, not all silhouettes can actuall y be projections of the same object in space: this simple observation has implic ations in object recognition and multi-view segmentation, and has been (often im plicitly) used as a basis for camera calibration. In this paper, we investigate the conditions for multiple silhouettes, or more generally arbitrary closed imag e sets, to be geometrically "consistent". We present this notion as a natural ge neralization of traditional multi-view geometry, which deals with consistency fo r points. After discussing some general results, we present a "dual" formulation for consistency, that gives conditions for a family of planar sets to be sectio ns of the same object. Finally, we introduce a more general notion of silhouette "compatibility" under partial knowledge of the camera projections, and point ou t some possible directions for future research.
****

Rolling Shutter Absolute Pose Problem With Known Vertical Direction
Cenek Albl, Zuzana Kukelova, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3355-3363
We present a solution to the rolling shutter (RS) absolute camera pose problem w ith known vertical direction. Our new solver, R5Pup, is an extension of the gene ral minimal solution R6P, which uses a double linearized RS camera model initial ized by the standard perspective P3P. Here, thanks to using known vertical direc tions, we avoid double linearization and can get the camera absolute pose direct ly from the RS model without the initialization by a standard P3P. Moreover, we need only five 2D-to-3D matches while R6P needed six such matches. We demonstrat e in simulated and real experiments that our new R5Pup is robust, fast and a ver

y practical method for absolute camera pose computation for modern cameras on mo
bile devices. We compare our R5Pup to the state of the art RS and perspective me
thods and demonstrate that it outperforms them when vertical direction is known
in the range of accuracy available on modern mobile devices. We also demonstrate
 that when using R5Pup solver in structure from motion (SfM) pipelines, it is be
tter to transform already reconstructed scenes into the standard position, rathe
r than using hard constraints on the verticality of up vectors.
****

Uncertainty-Driven 6D Pose Estimation of Objects and Scenes From a Single RGB Im
age
Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold
, carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2016, pp. 3364-3372
In recent years, the task of estimating the 6D pose of object instances and comp
lete scenes, i.e. camera localization, from a single input image has received co
nsiderable attention. Consumer RGB-D cameras have made this feasible, even for d
ifficult, texture-less objects and scenes. In this work, we show that a single R
GB image is sufficient to achieve visually convincing results. Our key concept i
s to model and exploit the uncertainty of the system at all stages of the proces
sing pipeline. The uncertainty comes in the form of continuous distributions ove
r 3D object coordinates and discrete distributions over object labels. We give t
hree technical contributions. Firstly, we develop a regularized, auto-context re
gression framework which iteratively reduces uncertainty in object coordinate an
d object label predictions. Secondly, we introduce an efficient way to marginali
ze object coordinate distributions over depth. This is necessary to deal with mi
ssing depth information. Thirdly, we utilize the distributions over object label
s to detect multiple objects simultaneously with a fixed budget of RANSAC hypoth
eses. We tested our system for object pose estimation and camera localization on
 commonly used data sets. We see a major improvement over competing systems.
****

Multicamera Calibration From Visible and Mirrored Epipoles
Andrey Bushnevskiy, Lorenzo Sorgi, Bodo Rosenhahn; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3373-3381
Multicamera rigs are used in a large number of 3D Vision applications, such as 3
D modeling, motion capture or telepresence and a robust calibration is of utmost
 importance in order to achieve a high accuracy results. In many practical confi
gurations the cameras in a rig are arranged in such a way, that they can observe
 each other, in other words a number of epipoles correspond to the real image po
ints. In this paper we propose a solution for the automatic recovery of the exte
rnal calibration of a multicamera system by enforcing only simple geometrical co
nstraints, arising from the epipole visibility, without using any calibration ob
ject, such as checkerboards, laser pointers or similar. Additionally, we introdu
ce an extension of the method that handles the case of epipoles being visible in
 the reflection of a planar mirror, which makes the algorithm suitable for the c
alibration of any multicamera system, irrespective of the number of cameras and
their actual mutual visibility, and furthermore we remark that it requires only
one or a few images per camera and therefore features a high speed and usability
. We produce an evidence of the algorithm effectiveness by presenting a wide set
 of tests performed on synthetic as well as real datasets and we compare the res
ults with those obtained using a traditional LED-based algorithm. The real datas
ets have been captured using a multicamera Virtual Reality (VR) rig and a spheri
cal dome configuration for 3D reconstruction.
****

Joint Unsupervised Deformable Spatio-Temporal Alignment of Sequences
Lazaros Zafeiriou, Epameinondas Antonakos, Stefanos Zafeiriou, Maja Pantic; Proc
eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
, 2016, pp. 3382-3390
Typically, the problems of spatial and temporal alignment of sequences are consi
dered disjoint. That is, in order to align two sequences, a methodology that (no
n)-rigidly aligns the images is first applied, followed by temporal alignment of

the obtained aligned images. In this paper, we propose the first, to the best of our knowledge, methodology that can jointly spatio-temporally align two sequences, which display highly deformable texture-varying objects. We show that by treating the problems of deformable spatial and temporal alignment jointly, we achieve better results than considering the problems independent. Furthermore, we show that deformable spatio-temporal alignment of faces can be performed in an unsupervised manner (i.e., without employing face trackers or building person-specific deformable models).
****

Deep Region and Multi-Label Learning for Facial Action Unit Detection
Kaili Zhao, Wen-Sheng Chu, Honggang Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3391-3399
Region learning (RL) and multi-label learning (ML) have recently attracted increasing attentions in the field of facial Action Unit (AU) detection. Knowing that AUs are active on sparse facial regions, RL aims to identify these regions for a better specificity. On the other hand, a strong statistical evidence of AU correlations suggests that ML is a natural way to model the detection task. In this paper, we propose Deep Region and Multi-label Learning (DRML), a unified deep network that simultaneously addresses these two problems. One crucial aspect in DRML is a novel region layer that uses feed-forward functions to induce important facial regions, forcing the learned weights to capture structural information of the face. Our region layer serves as an alternative design between locally connected layers (i.e., confined kernels to individual pixels) and conventional convolution layers (i.e., shared kernels across an entire image). Unlike previous studies that solve RL and ML alternately, DRML by construction addresses both problems, allowing the two seemingly irrelevant problems to interact more directly. The complete network is end-to-end trainable, and automatically learns representations robust to variations inherent within a local region. Experiments on BP4D and DISFA benchmarks show that DRML performs the highest average F1-score and AUC within and across datasets in comparison with alternative methods.
****

Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection
Yue Wu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3400-3408
Cascade regression framework has been shown to be effective for facial landmark detection. It starts from an initial face shape and gradually predicts the face shape update from the local appearance features to generate the facial landmark locations in the next iteration until convergence. In this paper, we improve upon the cascade regression framework and propose the Constrained Joint Cascade Regression Framework (CJCRF) for simultaneous facial action unit recognition and facial landmark detection, which are two related face analysis tasks, but are seldomly exploited together. In particular, we first learn the relationships among facial action units and face shapes as a constraint. Then, in the proposed constrained joint cascade regression framework, with the help from the constraint, we iteratively update the facial landmark locations and the action unit activation probabilities until convergence. Experimental results demonstrate that the intertwined relationships of facial action units and face shapes boost the performances of both facial action unit recognition and facial landmark detection. The experimental results also demonstrate the effectiveness of the proposed method comparing to the state-of-the-art works.
****

Unconstrained Face Alignment via Cascaded Compositional Learning
Shizhan Zhu, Cheng Li, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3409-3417
We present a practical approach to address the problem of unconstrained face alignment for a single image. In our unconstrained problem, we need to deal with large shape and appearance variations under extreme head poses and rich shape deformation. To equip cascaded regressors with the capability to handle global shape variation and irregular appearance-shape relation in the unconstrained scenario

, we partition the optimisation space into multiple domains of homogeneous desce
nt, and predict a shape as a composition of estimations from multiple domain-spe
cific regressors. With a specially formulated learning objective and a novel tre
e splitting function, our approach is capable of estimating a robust and meaning
ful composition. In addition to achieving state-of-the-art accuracy over existin
g approaches, our framework is also an efficient solution (350 FPS), thanks to t
he on-the-fly domain exclusion mechanism and the capability of leveraging the fa
st pixel feature.
****

Automated 3D Face Reconstruction From Multiple Images Using Quality Measures
Marcel Piotraschke, Volker Blanz; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2016, pp. 3418-3427
Automated 3D reconstruction of faces from images is challenging if the image mat
erial is difficult in terms of pose, lighting, occlusions and facial expressions
, and if the initial 2D feature positions are inaccurate or unreliable. We propo
se a method that reconstructs individual 3D shapes from multiple single images o
f one person, judges their quality and then combines the best of all results. Th
is is done separately for different regions of the face. The core element of thi
s algorithm and the focus of our paper is a quality measure that judges a recons
truction without information about the true shape. We evaluate different quality
 measures, develop a method for combining results, and present a complete proces
sing pipeline for automated reconstruction.
****

Occlusion-Free Face Alignment: Deep Regression Networks Coupled With De-Corrupt
AutoEncoders
Jie Zhang, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3428-3437
Face alignment or facial landmark detection plays an important role in many comp
uter vision applications, e.g., face recognition, facial expression recognition,
 face animation, etc. However, the performance of face alignment system degenera
tes severely when occlusions occur. In this work, we propose a novel face alignm
ent method, which cascades several Deep Regression networks coupled with De-corr
upt Autoencoders (denoted as DRDA) to explicitly handle partial occlusion proble
m. Different from the previous works that can only detect occlusions and discard
 the occluded parts, our proposed de-corrupt autoencoder network can automatical
ly recover the genuine appearance for the occluded parts and the recovered parts
 can be leveraged together with those non-occluded parts for more accurate align
ment. By coupling de-corrupt autoencoders with deep regression networks, a deep
alignment model robust to partial occlusions is achieved. Besides, our method ca
n localize occluded regions rather than merely predict whether the landmarks are
 occluded. Experiments on two challenging occluded face datasets demonstrate tha
t our method significantly outperforms the state-of-the-art methods.
****

Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis
Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Ca
navan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, Li
jun Yin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recog
nition (CVPR), 2016, pp. 3438-3446
Emotion is expressed in multiple modalities, yet most research has considered at
 most one or two. This stems in part from the lack of large, diverse, well-annot
ated, multimodal databases with which to develop and test algorithms. We present
 a well-annotated, multimodal, multidimensional spontaneous emotion corpus of 14
0 participants. Emotion inductions were highly varied. Data were acquired from a
 variety of sensors of the face that included high-resolution 3D dynamic imaging
, high-resolution 2D video, and thermal (infrared) sensing, and contact physiolo
gical sensors that included electrical conductivity of the skin, respiration, bl
ood pressure, and heart rate. Facial expression was annotated for both the occur
rence and intensity of facial action units from 2D video by experts in the Facia
l Action Coding System (FACS). The corpus further includes derived features from
 3D, 2D, and IR (infrared) sensors and baseline results for facial expression an

d action unit detection. The entire corpus will be made available to the researc
h community.
****

Learning Reconstruction-Based Remote Gaze Estimation
Pei Yu, Jiahuan Zhou, Ying Wu; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2016, pp. 3447-3455
It is a challenging problem to accurately estimate gazes from low-resolution eye
 images that do not provide fine and detailed features for eyes. Existing method
s attempt to establish the mapping between the visual appearance space to the ga
ze space. Different from the direct regression approach, the reconstruction-base
d approach represents appearance and gaze via local linear reconstruction in the
ir own spaces. A common treatment is to use the same local reconstruction in the
 two spaces, i.e., the reconstruction weights in the appearance space are transf
erred to the gaze space for gaze reconstruction. However, this questionable trea
tment is taken for granted but has never been justified, leading to significant
errors in gaze estimation. This paper is focused on the study of this fundamenta
l issue. It shows that the distance metric in the appearance space needs to be a
djusted, before the same reconstruction can be used. A novel method is proposed
to learn the metric, such that the affinity structure of the appearance space un
der this new metric is as close as possible to the affinity structure of the gaz
e space under the normal Euclidean metric. Furthermore, the local affinity struc
ture invariance is utilized to further regularize the solution to the reconstruc
tion weights, so as to obtain a more robust and accurate solution. Effectiveness
 of the proposed method is validated and demonstrated through extensive experime
nts on different subjects.
****

Joint Training of Cascaded CNN for Face Detection
Hongwei Qin, Junjie Yan, Xiu Li, Xiaolin Hu; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3456-3465
Cascade has been widely used in face detection, where classifier with low comput
ation cost can be firstly used to shrink most of the background while keeping th
e recall. The cascade in detection is popularized by seminal Viola-Jones framewo
rk and then widely used in other pipelines, such as DPM and CNN. However, to our
 best knowledge, most of the previous detection methods use cascade in a greedy
manner, where previous stages in cascade are fixed when training a new stage. So
 optimizations of different CNNs are isolated. In this paper, we propose joint t
raining to achieve end-to-end optimization for CNN cascade. We show that the bac
k propagation algorithm used in training CNN can be naturally used in training C
NN cascade. We present how jointly training can be conducted on naive CNN cascad
e and more sophisticated region proposal network (RPN) and fast R-CNN. Experimen
ts on face detection benchmarks verify the advantages of the joint training.
****

Facial Expression Intensity Estimation Using Ordinal Information
Rui Zhao, Quan Gan, Shangfei Wang, Qiang Ji; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3466-3474
Previous studies on facial expression analysis have been focused on recognizing
basic expression categories. There is limited amount of work on the continuous e
xpression intensity estimation, which is important for detecting and tracking em
otion change. Part of the reason is the lack of labeled data with annotated expr
ession intensity since expression intensity annotation requires expertise and is
 time consuming. In this work, we treat the expression intensity estimation as a
 regression problem. By taking advantage of the natural onset-apex-offset evolut
ion pattern of facial expression, the proposed method can handle different amoun
ts of annotations to perform frame-level expression intensity estimation. In ful
ly supervised case, all the frames are provided with intensity annotations. In w
eakly supervised case, only the annotations of selected key frames are used. Whi
le in unsupervised case, expression intensity can be estimated without any annot
ations. An efficient optimization algorithm based on Alternating Direction Metho
d of Multipliers (ADMM) is developed for solving the optimization problem associ
ated with parameter learning. We demonstrate the effectiveness of proposed metho

d by comparing it against both fully supervised and unsupervised approaches on b
enchmark facial expression datasets.
****

Proposal Flow
Bumsub Ham, Minsu Cho, Cordelia Schmid, Jean Ponce; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3475-3484
Finding image correspondences remains a challenging problem in the presence of i
ntra-class variations and large changes in scene layout. Semantic flow methods a
re designed to handle images depicting different instances of the same object or
 scene category. We introduce a novel approach to semantic flow, dubbed proposa
l flow, that establishes reliable correspondences using object proposals. Unlike
 prevailing semantic flow approaches that operate on pixels or regularly sampled
 local regions, proposal flow benefits from the characteristics of modern object
 proposals, that exhibit high repeatability at multiple scales, and can take adv
antage of both local and geometric consistency constraints among proposals. We a
lso show that proposal flow can effectively be transformed into a conventional d
ense flow field. We introduce a new dataset that can be used to evaluate both ge
neral semantic flow techniques and region-based approaches such as proposal flow
. We use this benchmark to compare different matching algorithms, object proposa
ls, and region features within proposal flow, to the state of the art in semanti
c flow. This comparison, along with experiments on standard datasets, demonstrat
es that proposal flow significantly outperforms existing semantic flow methods i
n various settings.
****

ProNet: Learning to Propose Object-Specific Boxes for Cascaded Neural Networks
Chen Sun, Manohar Paluri, Ronan Collobert, Ram Nevatia, Lubomir Bourdev; Proceed
ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2
016, pp. 3485-3493
This paper aims to classify and locate objects accurately and efficiently, witho
ut using bounding box annotations. It is challenging as objects in the wild coul
d appear at arbitrary locations and in different scales. In this paper, we propo
se a novel classification architecture ProNet based on convolutional neural netw
orks. It uses computationally efficient neural networks to propose image regions
 that are likely to contain objects, and applies more powerful but slower networ
ks on the proposed regions. The basic building block is a multi-scale fully-conv
olutional network which assigns object confidence scores to boxes at different l
ocations and scales. We show that such networks can be trained effectively using
 image-level annotations, and can be connected into cascades or trees for effici
ent object classification. ProNet outperforms previous state-of-the-art signific
antly on PASCAL VOC 2012 and MS COCO datasets for object classification and poin
t-based localization.
****

Seeing Behind the Camera: Identifying the Authorship of a Photograph
Christopher Thomas, Adriana Kovashka; Proceedings of the IEEE Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2016, pp. 3494-3502
We introduce the novel problem of identifying the photographer behind a photogra
ph. To explore the feasibility of current computer vision techniques to address
this problem, we created a new dataset of over 180,000 images taken by 41 well-k
nown photographers. Using this dataset, we examined the effectiveness of a varie
ty of features (low and high-level, including CNN features) at identifying the p
hotographer. We also trained a new deep convolutional neural network for this ta
sk. Our results show that high-level features greatly outperform low-level featu
res. We provide qualitative results using these learned models that give insight
 into our method's ability to distinguish between photographers, and allow us to
 draw interesting conclusions about what specific photographers shoot. We also d
emonstrate two applications of our method.
****

Material Classification Using Raw Time-Of-Flight Measurements
Shuochen Su, Felix Heide, Robin Swanson, Jonathan Klein, Clara Callenberg, Matth
ias Hullin, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vi

sion and Pattern Recognition (CVPR), 2016, pp. 3503-3511
We propose a material classification method using raw time-of-flight (ToF) measu rements. ToF cameras capture the correlation between a reference signal and the temporal response of material to incident illumination. Such measurements encode unique signatures of the material, i.e. the degree of subsurface scattering ins ide a volume. Subsequently, it offers an orthogonal domain of feature representa tion compared to conventional spatial and angular reflectance-based approaches. We demonstrate the effectiveness, robustness, and efficiency of our method throu gh experiments and comparisons of real-world materials.
****

Weakly Supervised Object Localization With Progressive Domain Adaptation
Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3512-3520
We address the problem of weakly supervised object localization where only image -level annotations are available for training. Many existing approaches tackle t his problem through object proposal mining. However, a substantial amount of noi se in object proposals causes ambiguities for learning discriminative object mod els. Such approaches are sensitive to model initialization and often converge to an undesirable local minimum. In this paper, we address this problem by progres sive domain adaptation with two main steps: classification adaptation and detect ion adaptation. In classification adaptation, we transfer a pre-trained network to our multi-label classification task for recognizing the presence of a certain object in an image. In detection adaptation, we first use a mask-out strategy t o collect class-specific object proposals and apply multiple instance learning t o mine confident candidates. We then use these selected object proposals to fine -tune all the layers, resulting in a fully adapted detection network. We extensi vely evaluate the localization performance on the PASCAL VOC and ILSVRC datasets and demonstrate significant performance improvement over the state-of-the-art m ethods.
****

Newtonian Scene Understanding: Unfolding the Dynamics of Objects in Static Image s
Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, Ali Farhadi; Proceed ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2 016, pp. 3521-3529
In this paper, we study the challenging problem of predicting the dynamics of ob jects in static images. Given a query object in an image, our goal is to provide a physical understanding of the object in terms of the forces acting upon it an d its long term motion as response to those forces. Direct and explicit estimati on of the forces and the motion of objects from a single image is extremely chal lenging. We define intermediate physical abstractions called Newtonian scenarios and introduce Newtonian Neural Network (N^3) that learns to map a single image to a state in a Newtonian scenario. Our experimental evaluations show that our m ethod can reliably predict dynamics of a query object from a single image. In ad dition, our approach can provide physical reasoning that supports the predicted dynamics in terms of velocity and force vectors. To spur research in this direct ion we compiled Visual Newtonian Dynamics (VIND) dataset that includes more than 6000 videos aligned with Newtonian scenarios represented using game engines, an d more than 4500 still images with their ground truth dynamics.
****

Identifying Good Training Data for Self-Supervised Free Space Estimation
Ali Harakeh, Daniel Asmar, Elie Shammas; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 3530-3538
This paper proposes a novel technique to extract training data from free space i n a scene using a stereo camera. The proposed technique exploits the projection of planes in the v-disparity image paired with Bayesian linear regression to rel iably identify training image pixels belonging to free space in a scene. Unlike other methods in the literature, the algorithm does not require any prior traini ng, has only one free parameter, and is shown to provide consistent results over

a variety of terrains without the need for any manual tuning. The proposed method is compared to two other data extraction methods from the literature. Results of Support Vector classifiers using training data extracted by the proposed technique are superior in terms of quality and consistency of free space estimation. Furthermore, the computation time required by the proposed technique is shown to be smaller and more consistent than that of other training data extraction methods.

****

Learning to Match Aerial Images With Deep Attentive Architectures

Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3539-3547

Image matching is a fundamental problem in Computer Vision. In the context of feature-based matching, SIFT and its variants have long excelled in a wide array of applications. However, for ultra-wide baselines, as in the case of aerial images captured under large camera rotations, the appearance variation goes beyond the reach of SIFT and RANSAC. In this paper we propose a data-driven, deep learning-based approach that sidesteps local correspondence by framing the problem as a classification task. Furthermore, we demonstrate that local correspondences can still be useful. To do so we incorporate an attention mechanism to produce a set of probable matches, which allows us to further increase performance. We train our models on a dataset of urban aerial imagery consisting of `same' and `different' pairs, collected for this purpose, and characterize the problem via a human study with annotations from Amazon Mechanical Turk. We demonstrate that our models outperform the state-of-the-art on ultra-wide baseline matching and approach human accuracy.

****

Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection

Krishna Kumar Singh, Fanyi Xiao, Yong Jae Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3548-3556

The status quo approach to training object detectors requires expensive bounding box annotations. Our framework takes a markedly different direction: we transfer tracked object boxes from weakly-labeled videos to weakly-labeled images to automatically generate pseudo ground-truth boxes, which replace manually annotated bounding boxes. We first mine discriminative regions in the weakly-labeled image collection that frequently/rarely appear in the positive/negative images. We then match those regions to videos and retrieve the corresponding tracked object boxes. Finally, we design a hough transform algorithm to vote for the best box to serve as the pseudo GT for each image, and use them to train an object detector. Together, these lead to state-of-the-art weakly-supervised detection results on the PASCAL 2007 and 2010 datasets.

****

DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns

Ali Diba, Ali Mohammad Pazandeh, Hamed Pirsiavash, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3557-3565

The recognition of human actions and the determination of human attributes are two tasks that call for fine-grained classification. Indeed, often rather small and inconspicuous objects and features have to be detected to tell their classes apart. In order to deal with this challenge, we propose a novel convolutional neural network that mines mid-level image patches that are sufficiently dedicated to resolve the corresponding subtleties. In particular, we train a newly designed CNN (DeepPattern) that learns discriminative patch groups. There are two innovative aspects to this. On the one hand we pay attention to contextual information in an original fashion. On the other hand, we let an iteration of feature learning and patch clustering purify the set of dedicated patches that we use. We validate our method for action classification on two challenging datasets: PASCAL VOC 2012 Action and Stanford 40 Actions, and for attribute recognition we use the Berkeley Attributes of People dataset. Our discriminative mid-level mi

ning CNN obtains state-of-the-art results on these datasets, without a need for annotations about parts and poses.
****

Canny Text Detector: Fast and Robust Scene Text Localization Algorithm
Hojin Cho, Myungchul Sung, Bongjin Jun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3566-3573
This paper presents a novel scene text detection algorithm, Canny Text Detector, which takes advantage of the similarity between image edge and text for effective text localization with improved recall rate. As closely related edge pixels construct the structural information of an object, we observe that cohesive characters compose a meaningful word/sentence sharing similar properties such as spatial location, size, color, and stroke width regardless of language. However, prevalent scene text detection approaches have not fully utilized such similarity, but mostly rely on the characters classified with high confidence, leading to low recall rate. By exploiting the similarity, our approach can quickly and robustly localize a variety of texts. Inspired by the original Canny edge detector, our algorithm makes use of double threshold and hysteresis tracking to detect texts of low confidence. Experimental results on public datasets demonstrate that our algorithm outperforms the state-of-the-art scene text detection methods in terms of detection rate.
****

Temporal Multimodal Learning in Audiovisual Speech Recognition
Di Hu, Xuelong Li, Xiaoqiang lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3574-3582
In view of the advantages of deep networks in producing useful representation, the generated features of different modality data (such as image, audio) can be jointly learned using Multimodal Restricted Boltzmann Machines (MRBM). Recently, audiovisual speech recognition based the MRBM has attracted much attention, and the MRBM shows its effectiveness in learning the joint representation across audiovisual modalities. However, the built networks have weakness in modeling the multimodal sequence which is the natural property of speech signal. In this paper, we will introduce a novel temporal multimodal deep learning architecture, named as Recurrent Temporal Multimodal RBM (RTMRBM), that models multimodal sequences by transforming the sequence of connected MRBMs into a probabilistic series model. Compared with existing multimodal networks, it's simple and efficient in learning temporal joint representation. We evaluate our model on audiovisual speech datasets, two public (AVLetters and AVLetters2) and one self-build. The experimental results demonstrate that our approach can obviously improve the accuracy of recognition compared with standard MRBM and the temporal model based on conditional RBM. In addition, RTMRBM still outperforms non-temporal multimodal deep networks in the presence of the weakness of long-term dependencies.
****

Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd
Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3583-3592
Object detection and 6D pose estimation in the crowd (scenes with multiple object instances, severe foreground occlusions and background distractors), has become an important problem in many rapidly evolving technological areas such as robotics and augmented reality. Single shot-based 6D pose estimators with manually designed features are still unable to tackle the above challenges, motivating the research towards unsupervised feature learning and next-best-view estimation. In this work, we present a complete framework for both single shot-based 6D object pose estimation and next-best-view prediction based on Hough Forests, the state of the art object pose estimator that performs classification and regression jointly. Rather than using manually designed features we a) propose an unsupervised feature learnt from depth-invariant patches using a Sparse Autoencoder and b) offer an extensive evaluation of various state of the art features. Furthermore, taking advantage of the clustering performed in the leaf nodes of Hough Forests, we learn to estimate the reduction of uncertainty in other views, formulating

the problem of selecting the next-best-view. To further improve pose estimation , we propose an improved joint registration and hypotheses verification module a s a final refinement step to reject false detections. We provide two additional challenging datasets inspired from realistic scenarios to extensively evaluate t he state of the art and our framework. One is related to domestic environments a nd the other depicts a bin-picking scenario mostly found in industrial settings. We show that our framework significantly outperforms state of the art both on p ublic and on our datasets.

****

Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to M ulti-View CNNs

Liuhao Ge, Hui Liang, Junsong Yuan, Daniel Thalmann; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3593-3601

Articulated hand pose estimation plays an important role in human-computer inter action. Despite the recent progress, the accuracy of existing methods is still n ot satisfactory, partially due to the difficulty of embedded high-dimensional an d non-linear regression problem. Different from the existing discriminative meth ods that regress for the hand pose with a single depth image, we propose to firs t project the query depth image onto three orthogonal planes and utilize these m ulti-view projections to regress for 2D heat-maps which estimate the joint posit ions on each plane. These multi-view heat-maps are then fused to produce final 3 D hand pose estimation with learned pose priors. Experiments show that the propo sed method largely outperforms state-of-the-arts on a challenging dataset. Moreo ver, a cross-dataset experiment also demonstrates the good generalization abilit y of the proposed method.

****

Semantic Segmentation With Boundary Neural Fields

Gedas Bertasius, Jianbo Shi, Lorenzo Torresani; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3602-3610

The state-of-the-art in semantic segmentation is currently represented by fully convolutional networks (FCNs). However, FCNs use large receptive fields and many pooling layers, both of which cause blurring and low spatial resolution in the deep layers. As a result FCNs tend to produce segmentations that are poorly loca lized around object boundaries. Prior work has attempted to address this issue i n post-processing steps, for example using a color-based CRF on top of the FCN p redictions. However, these approaches require additional parameters and low-leve l features that are difficult to tune and integrate into the original network ar chitecture. Additionally, most CRFs use color-based pixel affinities, which are not well suited for semantic segmentation and lead to spatially disjoint predict ions. To overcome these problems, we introduce a Boundary Neural Field (BNF), w hich is a global energy model integrating FCN predictions with boundary cues. Th e boundary information is used to enhance semantic segment coherence and to impr ove object localization. Specifically, we first show that the convolutional filt ers of semantic FCNs provide good features for boundary detection. We then emplo y the predicted boundaries to define pairwise potentials in our energy. Finally, we show that our energy decomposes semantic segmentation into multiple binary p roblems, which can be relaxed for efficient global optimization. We report exten sive experiments demonstrating that minimization of our global boundary-based en ergy yields results superior to prior globalization methods, both quantitatively as well as qualitatively.

****

HD Maps: Fine-Grained Road Segmentation by Parsing Ground and Aerial Images

Gellert Mattyus, Shenlong Wang, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 36 11-3619

In this paper we present an approach to enhance existing maps with fine grained segmentation categories such as parking spots and sidewalk, as well as the numbe r and location of road lanes. Towards this goal, we propose an efficient approac h that is able to estimate these fine grained categories by doing joint inferenc e over both, monocular aerial imagery, as well as ground images taken from a ste

reo camera pair mounted on top of a car. Important to this is reasoning about th e alignment between the two types of imagery, as even when the measurements are taken with sophisticated GPS+IMU systems, this alignment is not sufficiently acc urate. We demonstrate the effectiveness of our approach on a new dataset which e nhances KITTI [8] with aerial images taken with a camera mounted on an airplane and flying around the city of Karlsruhe, Germany.
****

DAG-Recurrent Neural Networks For Scene Labeling
Bing Shuai, Zhen Zuo, Bing Wang, Gang Wang; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3620-3629
In image labeling, local representations for image units (pixels, patches or sup erpixels) are usually generated from their surrounding image patches, thus long- range contextual information is not effectively encoded. In this paper, we intro duce recurrent neural networks (RNNs) to address this issue. Specifically, direc ted acyclic graph RNNs (DAG-RNNs) are proposed to process DAG-structured images, which enables the network to model long-range semantic dependencies among image units. Our DAG-RNNs are capable of tremendously enhancing the discriminative po wer of local representations, which significantly benefits the local classificat ion. Meanwhile, we propose a novel class weighting function that attends to rare classes, which phenomenally boosts the recognition accuracy for non-frequent cl asses. Integrating with convolution and deconvolution layers, our DAG-RNNs achie ve new state-of-the-art results on the challenging SiftFlow, CamVid and Barcelon a benchmarks.
****

Saliency Guided Dictionary Learning for Weakly-Supervised Image Parsing
Baisheng Lai, Xiaojin Gong; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2016, pp. 3630-3639
In this paper, we propose a novel method to perform weakly-supervised image pars ing based on the dictionary learning framework. To deal with the challenges caus ed by the label ambiguities, we design a saliency guided weight assignment schem e to boost the discriminative dictionary learning. More specifically, with a col lection of tagged images, the proposed method first conducts saliency detection and automatically infers the confidence for each semantic class to be foreground or background. These clues are then incorporated to learn the dictionaries, the weights, as well as the sparse representation coefficients in the meanwhile. On ce obtained the coefficients of a superpixel, we use a sparse representation cla ssifier to determine its semantic label. The approach is validated on the MSRC21 , PASCAL VOC07, and VOC12 datasets. Experimental results demonstrate the encoura ging performance of our approach in comparison with some state-of-the-arts.
****

Attention to Scale: Scale-Aware Semantic Image Segmentation
Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, Alan L. Yuille; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3 640-3649
Incorporating multi-scale features in fully convolutional neural networks (FCNs) has been a key element to achieving state-of-the-art performance on semantic im age segmentation. One common way to extract multi-scale features is to feed mult iple resized input images to a shared deep network and then merge the resulting features for pixel-wise classification. In this work, we propose an attention me chanism that learns to softly weight the multi-scale features at each pixel loca tion. We adapt a state-of-the-art semantic image segmentation model, which we jo intly train with multi-scale input images and the attention model. The proposed attention model not only outperforms average- and max-pooling, but allows us to diagnostically visualize the importance of features at different positions and s cales. Moreover, we show that adding extra supervision to the output at each sca le is essential to achieving excellent performance when merging multi-scale feat ures. We demonstrate the effectiveness of our model with extensive experiments o n three challenging datasets, including PASCAL-Person-Part, PASCAL VOC 2012 and a subset of MS-COCO 2014.
****

Scene Labeling Using Sparse Precision Matrix
Nasim Souly, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3650-3658
Scene labeling task is to segment the image into meaningful regions and categorize them into classes of objects which comprised the image. Commonly used methods typically find the local features for each segment and label them using classifiers. Afterwards, labeling is smoothed in order to make sure that neighboring regions receive similar labels. However, these methods ignore expressive connections between labels and non-local dependencies among regions. In this paper, we propose to use a sparse estimation of precision matrix (also called concentration matrix), which is the inverse of covariance matrix of data obtained by graphical lasso to find interaction between labels and regions. To do this, we formulate the problem as an energy minimization over a graph, whose structure is captured by applying sparse constraint on the elements of the precision matrix. This graph encodes (or represents) only significant interactions and avoids a fully connected graph, which is typically used to reflect the long distance associations. We use local and global information to achieve better labeling. We assess our approach on three datasets and obtained promising results.
****

Iterative Instance Segmentation
Ke Li, Bharath Hariharan, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3659-3667
Existing methods for pixel-wise labelling tasks generally disregard the underlying structure of labellings, often leading to predictions that are visually implausible. While incorporating structure into the model should improve prediction quality, doing so is challenging - manually specifying the form of structural constraints may be impractical and inference often becomes intractable even if structural constraints are given. We sidestep this problem by reducing structured prediction to a sequence of unconstrained prediction problems and demonstrate that this approach is capable of automatically discovering priors on shape, contiguity of region predictions and smoothness of region contours from data without any a priori specification. On the instance segmentation task, this method outperforms the state-of-the-art, achieving a mean AP^r of 63.6% at 50% overlap and 43.3% at 70% overlap.
****

Recurrent Attentional Networks for Saliency Detection
Jason Kuen, Zhenhua Wang, Gang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3668-3677
Convolutional-deconvolution networks can be adopted to perform end-to-end saliency detection. But, they do not work well with objects of multiple scales. To overcome such a limitation, in this work, we propose a recurrent attentional convolutional-deconvolution network (RACDNN). Using spatial transformer and recurrent network units, RACDNN is able to iteratively attend to selected image sub-regions to perform saliency refinement progressively. Besides tackling the scale problem, RACDNN can also learn context-aware features from past iterations to enhance saliency refinement in future iterations. Experiments on several challenging saliency detection datasets validate the effectiveness of RACDNN, and show that RACDNN outperforms state-of-the-art saliency detection methods.
****

Instance-Level Video Segmentation From Object Tracks
Guillaume Seguin, Piotr Bojanowski, Remi Lajugie, Ivan Laptev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3678-3687
We address the problem of segmenting multiple object instances in complex videos. Our method does not require manual pixel-level annotation for training, and relies instead on readily-available object detectors or visual object tracking only. Given object bounding boxes at input, we cast video segmentation as a weakly-supervised learning problem. Our proposed objective combines (a) a discriminative clustering term for background segmentation, (b) a spectral clustering one for grouping pixels of same object instances, and (c) linear constraints enabling i

nstance-level segmentation. We propose a convex relaxation of this problem and s
olve it efficiently using the Frank-Wolfe algorithm. We report results and compa
re our method to several baselines on a new video dataset for multi-instance per
son segmentation.
****

Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer
Jun Xie, Martin Kiefel, Ming-Ting Sun, Andreas Geiger; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3688-3697
This supplementary material provides additional illustrations, visualizations an
d experiments. We start by showing the color coding and label mapping used for t
he semantic and instance label results in the paper. Then we provide more detail
s about the 3D fold/curb detection and parameter settings that are used in the p
aper. Next, we provide additional quantitative and qualitative semi-dense infere
nce results for both semantic and instance segmentation. Finally, we show the ab
ility of our method to annotate 3D point clouds with semantic and instance label
s which is a byproduct of our approach.
****

Amplitude Modulated Video Camera - Light Separation in Dynamic Scenes
Amir Kolaman, Maxim Lvov, Rami Hagege, Hugo Guterman; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3698-3706
Controlled light conditions improve considerably the performance of most compute
r vision algorithms. Dynamic light conditions create varying spatial changes in
color and intensity across the scene. These condition, caused by a moving shadow
 for example, force developers to create algorithms which are robust to such var
iations. We suggest a computational camera which produces images that are not in
fluenced by environmental variations in light conditions. The key insight is tha
t many years ago, similar difficulties were already solved in radio communicatio
n; As a result each channel is immune to interference from other radio channels.
 Amplitude Modulated (AM) video camera separates the influence of a modulated l
ight from other unknown light sources in the scene; Causing the AM video camera
frame to appear the same - independent of the light conditions in which it was
taken. We built a prototype of the AM video camera by using off the shelf hardwa
re and tested it. AM video camera was used to demonstrate color constancy, shado
w removal and contrast enhancement in real time. We show theoretically and empir
ically that: 1. the proposed system can produce images with similar noise levels
 as a standard camera. 2. The images created by such camera are almost completel
y immune to temporal, spatial and spectral changes in the background light.
****

A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometr
ic Stereo
Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, Ping Tan; Proceedin
gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201
6, pp. 3707-3716
Recent progress on photometric stereo extends the technique to deal with general
 materials and unknown illumination conditions. However, due to the lack of suit
able benchmark data with ground truth shapes (normals), quantitative comparison
and evaluation is difficult to achieve. In this paper, we first survey and categ
orize existing methods using a photometric stereo taxonomy emphasizing on non-La
mbertian and uncalibrated methods. We then introduce the 'DiLiGenT' photometric
stereo image dataset with calibrated Directional Lightings, objects of General r
eflectance, and 'ground Truth' shapes (normals). Based on our dataset, we quanti
tatively evaluate state-of-the-art photometric stereo methods for general non-La
mbertian materials and unknown lightings to analyze their strengths and limitati
ons.
****

Depth From Semi-Calibrated Stereo and Defocus
Ting-Chun Wang, Manohar Srikanth, Ravi Ramamoorthi; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3717-3726
In this work, we propose a multi-camera system where we combine a main high-qual
ity camera with two low-res auxiliary cameras. The auxiliary cameras are well ca

librated and act as a passive depth sensor by generating disparity maps. The mai
n camera has an interchangeable lens and can produce good quality images at high
 resolution. Our goal is, given the low-res depth map from the auxiliary cameras
, generate a depth map from the viewpoint of the main camera. The advantage of o
ur system, compared to other systems such as light-field cameras or RGBD sensors
, is the ability to generate a high-resolution color image with a complete depth
 map, without sacrificing resolution and with minimal auxiliary hardware. Since
 the main camera has an interchangeable lens, it cannot be calibrated beforehand
, and directly applying stereo matching on it and either of the auxiliary camera
s often leads to unsatisfactory results. Utilizing both the calibrated cameras a
t once, we propose a novel approach to better estimate the disparity map of the
main camera. Then by combining the defocus cue of the main camera, the disparity
 map can be further improved. We demonstrate the performance of our algorithm on
 various scenes.
****

Exploiting Spectral-Spatial Correlation for Coded Hyperspectral Image Restoratio
n

Ying Fu, Yinqiang Zheng, Imari Sato, Yoichi Sato; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3727-3736

Conventional scanning and multiplexing techniques for hyperspectral imaging suff
er from limited temporal and/or spatial resolution. To resolve this issue, codin
g techniques are becoming increasingly popular in developing   snapshot systems
for high-resolution hyperspectral imaging. For such systems, it is   a critical
 task to accurately restore the 3D hyperspectral image   from its corresponding
 coded 2D image. In this paper, we propose an   effective method for coded hype
rspectral image restoration, which exploits extensive structure sparsity in the
hyperspectral image. Specifically, we simultaneously explore spectral and spatia
l correlation via low-rank regularizations, and formulate the restoration proble
m into a variational optimization model, which can be solved via an iterative nu
merical algorithm. Experimental   results using both synthetic data and real ima
ges show that the proposed method can significantly outperform the   state-of-t
he-art methods on several popular coding-based hyperspectral imaging systems.
****

Variable Aperture Light Field Photography: Overcoming the Diffraction-Limited Sp
atio-Angular Resolution Tradeoff

Julie Chang, Isaac Kauvar, Xuemei Hu, Gordon Wetzstein; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3737-374
5

Light fields have many applications in machine vision, consumer photography, rob
otics, and microscopy. However, the prevalent resolution limits of existing ligh
t field imaging systems hinder widespread adoption. In this paper, we analyze fu
ndamental resolution limits of light field cameras in the diffraction limit. We
propose a sequential, coded-aperture-style acquisition scheme that optimizes the
 resolution of a light field reconstructed from multiple photographs captured fr
om different perspectives and f-number settings. We also show that the proposed
acquisition scheme facilitates high dynamic range light field imaging and demons
trate a proof-of-concept prototype system. With this work, we hope to advance ou
r understanding of the resolution limits of light field photography and develop
practical computational imaging systems to overcome them.
****

Convolutional Networks for Shape From Light Field

Stefan Heber, Thomas Pock; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2016, pp. 3746-3754

Convolutional Neural Networks (CNNs) have recently been successfully applied to
various Computer Vision (CV) applications. In this paper we utilize CNNs to pred
ict depth information for given Light Field (LF) data. The proposed method learn
s an end-to-end mapping between the 4D light field and a representation of the c
orresponding 4D depth field in terms of 2D hyperplane orientations. The obtained
 prediction is then further refined in a post processing step by applying a high
er-order regularization. Existing LF datasets are not sufficient for the purpose

of the training scheme tackled in this paper. This is mainly due to the fact th
at the ground truth depth of existing datasets is inaccurate and/or the datasets
are limited to a small number of LFs. This made it necessary to generate a new
synthetic LF dataset, which is based on the raytracing software POV-Ray. This ne
w dataset provides floating point accurate ground truth depth fields, and due to
a random scene generator the dataset can be scaled as required.
****

Panoramic Stereo Videos With a Single Camera
Rajat Aggarwal, Amrisha Vohra, Anoop M. Namboodiri; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3755-3763
We present a practical solution for generating 360 degree stereo panoramic video
s using a single camera. Current approaches either use a moving camera that capt
ures multiple images of a scene, which are then stitched together to form the fi
nal panorama, or use multiple cameras that are synchronized. A moving camera lim
its the solution to static scenes, while multi-camera solutions require dedicate
d calibrated setups. Our approach improves upon the existing solutions in two si
gnificant ways: It solves the problem using a single camera, thus minimizing the
calibration problem and providing us the ability to convert any digital camera
into a panoramic stereo capture device. It captures all the light rays required
for stereo panoramas in a single frame using a compact custom designed mirror, t
hus making the design practical to manufacture and easier to use. We analyze sev
eral properties of the design as well as present panoramic stereo and depth esti
mation results.
****

The Next Best Underwater View
Mark Sheinin, Yoav Y. Schechner; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2016, pp. 3764-3773
To image in high resolution large and occlusion-prone scenes, a camera must move
above and around. Degradation of visibility due to geometric occlusions and dis
tances is exacerbated by scattering, when the scene is in a participating medium
. Moreover, underwater and in other media, artificial lighting is needed. Overal
l, data quality depends on the observed surface, medium and the time-varying pos
es of the camera and light source. This work proposes to optimize camera and lig
ht poses as they move, so that the surface is scanned efficiently and the descat
tered recovery has the highest quality. The work generalizes the next best view
concept of robot vision to scattering media and cooperative movable lighting. It
also extends descattering to platforms that move optimally. The optimization cr
iterion is information gain, taken from information theory. We exploit the exist
ence of a prior rough 3D model, since underwater such a model is routinely obtai
ned using sonar. We demonstrate this principle in a scaled-down setup.
****

Reconstructing Shapes and Appearances of Thin Film Objects Using RGB Images
Yoshie Kobayashi, Tetsuro Morimoto, Imari Sato, Yasuhiro Mukaigawa, Takao Tomono
, Katsushi Ikeuchi; Proceedings of the IEEE Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2016, pp. 3774-3782
Reconstruction of shapes and appearances of thin film objects can be applied to
many fields such as industrial inspection, biological analysis, and archeology r
esearch. However, it comes with many challenging issues because the appearances
of thin film can change dramatically depending on view and light directions. The
appearance is deeply dependent on not only the shapes but also the optical para
meters of thin film. In this paper, we propose a novel method to estimate shapes
and film thickness. First, we narrow down candidates of zenith angle by degree
of polarization and determine it by the intensity of thin film which increases m
onotonically along the zenith angle. Second, we determine azimuth angle from occ
luding boundaries. Finally, we estimate the film thickness by comparing a look-u
p table of color along the thickness and zenith angle with captured images. We e
xperimentally evaluated the accuracy of estimated shapes and appearances and fou
nd that our proposed method is effective.
****

Noisy Label Recovery for Shadow Detection in Unfamiliar Domains

Tomas F. Yago Vicente, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3783-3792
Recent shadow detection algorithms have shown initial success on small datasets of images from specific domains. However, shadow detection on broader image domains is still challenging due to the lack of annotated training data. This is due to the intense manual labor in annotating shadow data. In this paper we propose "lazy annotation", an efficient annotation method where an annotator only needs to mark the important shadow areas and some non-shadow areas. This yields data with noisy labels that are not yet useful for training a shadow detector. We address the problem of label noise by jointly learning a shadow region classifier and recovering the labels in the training set. We consider the training labels as unknowns and formulate the label recovery problem as the minimization of the sum of squared leave-one-out errors of a Least Squares SVM, which can be efficiently optimized. Experimental results show that a classifier trained with recovered labels achieves comparable performance to a classifier trained on the properly annotated data. These results suggest a feasible approach to address the task of detecting shadows in an unfamiliar domain: collecting and lazily annotating some images from the new domain for training. As will be demonstrated, this approach outperforms methods that rely on precisely annotated but less relevant datasets. Initial results suggest more general applicability.
****

Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled
Oscar Koller, Hermann Ney, Richard Bowden; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3793-3802
This work presents a new approach to learning a frame-based classifier on weakly labelled sequence data by embedding a CNN within an iterative EM algorithm. This allows the CNN to be trained on a vast number of example images when only loose sequence level information is available for the source videos. Although we demonstrate this in the context of hand shape recognition, the approach has wider application to any video recognition task where frame level labelling is not available. The iterative EM algorithm leverages the discriminative ability of the CNN to iteratively refine the frame level annotation and subsequent training of the CNN. By embedding the classifier within an EM framework the CNN can easily be trained on 1 million hand images. We demonstrate that the final classifier generalises over both individuals and data sets. The algorithm is evaluated on over 3000 manually labelled hand shape images of 60 different classes which will be released to the community. Furthermore, we demonstrate its use in continuous sign language recognition on two publicly available large sign language data sets, where it outperforms the current state-of-the-art by a large margin. To our knowledge no previous work has explored expectation maximization without Gaussian mixture models to exploit weak sequence labels for sign language recognition.
****

Recognizing Car Fluents From Video
Bo Li, Tianfu Wu, Caiming Xiong, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3803-3812
Physical fluents, a term originally used by Newton [40], refers to time-varying object states in dynamic scenes. In this paper, we are interested in inferring the fluents of vehicles from video. For example, a door (hood, trunk) is open or closed through various actions, light is blinking to turn. Recognizing these fluents has broad applications, yet have received scant attention in the computer vision literature. Car fluent recognition entails a unified framework for car detection, car part localization and part status recognition, which is made difficult by large structural and appearance variations, low resolutions and occlusions. This paper learns a spatial-temporal And-Or hierarchical model to represent car fluents. The learning of this model is formulated under the latent structural SVM framework. Since there are no publicly related dataset, we collect and annotate a car fluent dataset consisting of car videos with diverse fluents. In experiments, the proposed method outperforms several highly related baseline methods in terms of car fluent recognition and car part localization.

****
Pairwise Decomposition of Image Sequences for Active Multi-View Recognition
Edward Johns, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3813-3822
A multi-view image sequence provides a much richer capacity for object recognition than from a single image. However, most existing solutions to multi-view recognition typically adopt hand-crafted, model-based geometric methods, which do not readily embrace recent trends in deep learning. We propose to bring Convolutional Neural Networks to generic multi-view recognition, by decomposing an image sequence into a set of image pairs, classifying each pair independently, and then learning an object classifier by weighting the contribution of each pair. This allows for recognition over arbitrary camera trajectories, without requiring explicit training over the potentially infinite number of camera paths and lengths. Building these pairwise relationships then naturally extends to the next-best-view problem in an active recognition framework. To achieve this, we train a second Convolutional Neural Network to map directly from an observed image to next viewpoint. Finally, we incorporate this into a trajectory optimisation task, whereby the best recognition confidence is sought for a given trajectory length. We present state-of-the-art results in both guided and unguided multi-view recognition on the ModelNet dataset, and show how our method can be used with depth images, greyscale images, or both.
****
Inferring Forces and Learning Human Utilities From Videos
Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3823-3833
We propose a notion of affordance that takes into account physical quantities generated when the human body interacts with real-world objects, and introduce a learning framework that incorporates the concept of human utilities, which in our opinion provides a deeper and finer-grained account not only of object affordance but also of people's interaction with objects. Rather than defining affordance in terms of the geometric compatibility between body poses and 3D objects, we devise algorithms that employ physics-based simulation to infer the relevant forces/pressures acting on body parts. By observing the choices people make in videos (particularly in selecting a chair in which to sit) our system learns the comfort intervals of the forces exerted on body parts (while sitting). We account for people's preferences in terms of human utilities, which transcend comfort intervals to account also for meaningful tasks within scenes and spatiotemporal constraints in motion planning, such as for the purposes of robot task planning.
****
Force From Motion: Decoding Physical Sensation in a First Person Video
Hyun Soo Park, jyh-Jing Hwang, Jianbo Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3834-3842
A first-person video can generate powerful physical sensations of action in an observer. In this paper, we focus on a problem of Force from Motion---decoding the sensation of 1) passive forces such as the gravity, 2) the physical scale of the motion (speed) and space, and 3) active forces exerted by the observer such as pedaling a bike or banking on a ski turn. The sensation of gravity can be observed in a natural image. We learn this image cue for predicting a gravity direction in a 2D image and integrate the prediction across images to estimate the 3D gravity direction using structure from motion. The sense of physical scale is revealed to us when the body is in a dynamically balanced state. We compute the unknown physical scale of 3D reconstructed camera motion by leveraging the torque equilibrium at a banked turn that relates the centripetal force, gravity, and the body leaning angle. The active force and torque governs 3D egomotion through the physics of rigid body dynamics. Using an inverse dynamics optimization, we directly minimize 2D reprojection error (in video) with respect to 3D world structure, active forces, and additional passive forces such as air drag and friction force. We use structure from motion with the physical scale and gravity direction as an initialization of our bundle adjustment for force estimation. Our met

hod shows quantitatively equivalent reconstruction comparing to IMU measurements in terms of gravity and scale recovery and outperforms method based on 2D optical flow for an active action recognition task. We apply our method to first person videos of mountain biking, urban bike racing, skiing, speedflying with parachute, and wingsuit flying where inertial measurements are not accessible.
****

Robust Multi-Body Feature Tracker: A Segmentation-Free Approach
Pan Ji, Hongdong Li, Mathieu Salzmann, Yiran Zhong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3843-3851
Feature tracking is a fundamental problem in computer vision with applications in various tasks including 3D reconstruction and visual SLAM. While many methods have been devoted to making these tasks robust to noise and outliers, less attention has been attracted to improving the feature tracking itself. This paper introduces a novel multi-body feature tracker that takes advantage of the multi-body rigidity assumption to improve tracking robustness. A conventional approach to addressing this problem would consist of alternating between solving two subtasks: motion segmentation and feature tracking under rigidity constraints for each segment. This approach, however, requires knowing the number of motions, as well as assigning points to motion groups, which is typically sensitive to the motion estimates. By contrast, here, we introduce a segmentation-free solution to multi-body feature tracking that bypasses the motion assignment step and reduces to solving a series of subproblems with closed-form solutions. Our experiments demonstrate the benefits of our approach in terms of tracking accuracy and robustness to noise.
****

Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video
Dinesh Jayaraman, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3852-3861
How can unlabeled video augment visual learning? Existing methods perform "slow" feature analysis, encouraging temporal coherence, where the image representations of temporally close frames to exhibit only small differences. While this standard approach captures the fact that high-level visual signals change slowly over time, it fails to capture *how* the visual content changes. We propose to generalize slow feature analysis to "steady" feature analysis. The key idea is to impose a prior that higher order derivatives in the learned feature space must be small. To this end, we train a convolutional neural network with a regularizer that minimizes a contrastive loss on tuples of sequential frames from unlabeled video. Focusing on the case of triplets of frames, the proposed method encourages that feature changes over time should be smooth, i.e., similar to the most recent changes. Using five diverse image and video datasets, including unlabeled YouTube and KITTI videos, we demonstrate our method's impact on object recognition, scene classification, and action recognition tasks. We further show that our features learned from unlabeled video can even surpass a standard heavily supervised pretraining approach.
****

Volumetric 3D Tracking by Detection
Chun-Hao Huang, Benjamin Allain, Jean-Sebastien Franco, Nassir Navab, Slobodan Ilic, Edmond Boyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3862-3870
In this paper, we propose a new framework for 3D tracking by detection based on fully volumetric representations. On one hand, 3D tracking by detection has shown robust use in the context of interaction (Kinect) and surface tracking. On the other hand, volumetric representations have recently been proven efficient both for building 3D features and for addressing the 3D tracking problem. We leverage these benefits by unifying both families of approaches into a single, fully volumetric tracking-by-detection framework. We use a centroidal Voronoi tessellation (CVT) representation to compactly tessellate shapes with optimal discretization, construct a feature space, and perform the tracking according to the correspondences provided by trained random forests. Our results show improved tracking and training computational efficiency and improved memory performance. This in t

urn enables the use of larger training databases than state of the art approaches, which we leverage by proposing a cross-tracking subject training scheme to benefit from all subject sequences for all tracking situations, thus yielding better detection and less overfitting.

****

The Solution Path Algorithm for Identity-Aware Multi-Object Tracking
Shoou-I Yu, Deyu Meng, Wangmeng Zuo, Alexander Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3871-3879

We propose an identity-aware multi-object tracker based on the solution path algorithm. Our tracker not only produces identity-coherent trajectories based on cues such as face recognition, but also has the ability to pinpoint potential tracking errors. The tracker is formulated as a quadratic optimization problem with L0 norm constraints, which we propose to solve with the solution path algorithm. The algorithm successively solves the same optimization problem but under different Lp norm constraints, where p gradually decreases from 1 to 0. Inspired by the success of the solution path algorithm in various machine learning tasks, this strategy is expected to converge to a better local minimum than directly minimizing the hardly solvable L0 norm or the roughly approximated L1 norm constraints. Furthermore, the acquired solution path complies with the "decision making process" of the tracker, which provides more insight to locating potential tracking errors. Experiments show that not only is our proposed tracker effective, but also the solution path enables automatic pinpointing of potential tracking failures, which can be readily utilized in an active learning framework to improve identity-aware multi-object tracking.

****

In Defense of Sparse Tracking: Circulant Sparse Tracker
Tianzhu Zhang, Adel Bibi, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3880-3888

Sparse representation has been introduced to visual tracking by finding the best target candidate with minimal reconstruction error within the particle filter framework. However, most sparse representation based trackers have high computational cost, less than promising tracking performance, and limited feature representation. To deal with the above issues, we propose a novel circulant sparse tracker (CST), which exploits circulant target templates. Because of the circulant structure property, CST has the following advantages: (1) It can refine and reduce particles using circular shifts of target templates. (2) The optimization can be efficiently solved entirely in the Fourier domain. (3) High dimensional features can be embedded into CST to significantly improve tracking performance without sacrificing much computation time. Both qualitative and quantitative evaluations on challenging benchmark sequences demonstrate that CST performs better than all other sparse trackers and favorably against state-of-the-art methods.

****

Optical Flow With Semantic Segmentation and Localized Layers
Laura Sevilla-Lara, Deqing Sun, Varun Jampani, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3889-3898

Existing optical flow methods make generic, spatially homogeneous, assumptions about the spatial structure of the flow. In reality, optical flow varies across an image depending on object class. Simply put, different objects move differently. Here we exploit recent advances in static semantic scene segmentation to segment the image into objects of different types. We define different models of image motion in these regions depending on the type of object. For example, the road motion with homographies, vegetation with spatially smooth flow, and independently moving objects like cars and planes with affine+deviations. We then pose the flow estimation problem using a novel formulation of localized layers, which addresses limitations of traditional layered models for dealing with complex scene motion. Our semantic flow method achieves the lowest error of any published method in the KITTI-2015 flow benchmark and produces qualitatively better flow and segmentation than recent top methods on a wide range of natural videos.

****
Video Segmentation via Object Flow
Yi-Hsuan Tsai, Ming-Hsuan Yang, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3899-3908
Video object segmentation is challenging due to fast moving objects, deforming shapes, and cluttered backgrounds. Optical flow can be used to propagate an object segmentation over time but, unfortunately, flow is often inaccurate, particularly around object boundaries. Such boundaries are precisely where we want our segmentation to be accurate. To obtain accurate segmentation across time, we propose an efficient algorithm that considers video segmentation and optical flow estimation simultaneously. For video segmentation, we formulate a principled, multi-scale, spatio-temporal objective function that uses optical flow to propagate information between frames. For optical flow estimation, particularly at object boundaries, we compute the flow independently in the segmented regions and recompose the results. We call the process object flow and demonstrate the effectiveness of jointly optimizing optical flow and video segmentation using an iterative scheme. Experiments on the SegTrack v2 and Youtube-Objects datasets show that the proposed algorithm performs favorably against the other state-of-the-art methods.
****
Closed-Form Training of Mahalanobis Distance for Supervised Clustering
Marc T. Law, YaoLiang Yu, Matthieu Cord, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3909-3917
Clustering is the task of grouping a set of objects so that objects in the same cluster are more similar to each other than to those in other clusters. The crucial step in most clustering algorithms is to find an appropriate similarity metric, which is both challenging and problem-dependent. Supervised clustering approaches, which can exploit labeled clustered training data that share a common metric with the test set, have thus been proposed. Unfortunately, current metric learning approaches for supervised clustering do not scale to large or even medium-sized datasets. In this paper, we propose a new structured Mahalanobis Distance Metric Learning method for supervised clustering. We formulate our problem as an instance of large margin structured prediction and prove that it can be solved very efficiently in closed-form. The complexity of our method is (in most cases) linear in the size of the training dataset. We further reveal a striking similarity between our approach and multivariate linear regression. Experiments on both synthetic and real datasets confirm several orders of magnitude speedup while still achieving state-of-the-art performance.
****
Scalable Sparse Subspace Clustering by Orthogonal Matching Pursuit
Chong You, Daniel Robinson, Rene Vidal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3918-3927
Subspace clustering methods based on ell_1, l_2 or nuclear norm regularization have become very popular due to their simplicity, theoretical guarantees and empirical success. However, the choice of the regularizer can greatly impact both theory and practice. For instance, ell_1 regularization is guaranteed to give a subspace-preserving affinity (i.e., there are no connections between points from different subspaces) under broad conditions e.g., arbitrary subspaces and corrupted data). However, it requires solving a large scale convex optimization problem. On the other hand, l_2 and nuclear norm regularization provide efficient closed form solutions, but require very strong assumptions to guarantee a subspace-preserving affinity, e.g., independent subspaces and uncorrupted data. In this paper we study a subspace clustering method based on orthogonal matching pursuit. We show that the method is both computationally efficient and guaranteed to give a subspace-preserving affinity under broad conditions. Experiments on synthetic data verify our theoretical analysis, and applications in handwritten digit and face clustering show that our approach achieves the best trade off between accuracy and efficiency. Moreover, our approach is the first one to handle 100,000 data points.
****

Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering
Chong You, Chun-Guang Li, Daniel P. Robinson, Rene Vidal; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3928-3 937
State-of-the-art subspace clustering methods are based on expressing each data p oint as a linear combination of other data points while regularizing the matrix of coefficients with l_1, l_2 or nuclear norms. l_1 regularization is guaranteed to give a subspace-preserving affinity (i.e., there are no connections between points from different subspaces) under broad theoretical conditions, but the clu sters may not be connected. l_2 and nuclear norm regularization often improve co nnectivity, but give a subspace-preserving affinity only for independent subspac es. Mixed l_1, l_2 and nuclear norm regularizations offer a balance between the subspace-preserving and connectedness properties, but this comes at the cost of increased computational complexity. This paper studies the geometry of the elast ic net regularizer (a mixture of the l_1 and l_2 norms) and uses it to derive a provably correct and scalable active set method for finding the optimal coeffici ents. Our geometric analysis also provides a theoretical justification and a geo metric interpretation for the balance between the connectedness (due to l_2 regu larization) and subspace-preserving (due to l_1 regularization) properties for e lastic net subspace clustering. Our experiments show that the proposed active se t method not only achieves state-of-the-art clustering performance, but also eff iciently handles large-scale datasets.
****

Sparse Coding and Dictionary Learning With Linear Dynamical Systems
Wenbing Huang, Fuchun Sun, Lele Cao, Deli Zhao, Huaping Liu, Mehrtash Harandi; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2016, pp. 3938-3947
Linear Dynamical Systems (LDSs) are the fundamental tools for encoding spatio-te mporal data in various disciplines. To enhance the performance of LDSs, in this paper, we address the challenging issue of performing sparse coding on the space of LDSs, where both data and dictionary atoms are LDSs. Rather than approximate the extended observability with a finite-order matrix, we represent the space o f LDSs by an infinite Grassmannian consisting of the orthonormalized extended ob servability subspaces. Via a homeomorphic mapping, such Grassmannian is embedded into the space of symmetric matrices, where a tractable objective function can be derived for sparse coding. Then, we propose an efficient method to learn the system parameters of the dictionary atoms explicitly, by imposing the symmetric constraint to the transition matrices of the data and dictionary systems. Moreov er, we combine the state covariance into the algorithm formulation, thus further promoting the performance of the models with symmetric transition matrices. Com parative experimental evaluations reveal the superior performance of proposed me thods on various tasks including video classification and tactile recognition.
****

Sublabel-Accurate Relaxation of Nonconvex Energies
Thomas Mollenhoff, Emanuel Laude, Michael Moeller, Jan Lellmann, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3948-3956
We propose a novel spatially continuous framework for convex relaxations based o n functional lifting. Our method can be interpreted as a sublabel-accurate solut ion to multilabel problems. We show that previously proposed functional lifting methods optimize an energy which is linear between two labels and hence require (often infinitely) many labels for a faithful approximation. In contrast, the pr oposed formulation is based on a piecewise convex approximation and therefore ne eds far fewer labels - see Fig. 1. In comparison to recent MRF-based approaches, our method is formulated in a spatially continuous setting and shows less grid bias. Moreover, in a local sense, our formulation is the tightest possible conve x relaxation. It is easy to implement and allows an efficient primal-dual optimi zation on GPUs. We show the effectiveness of our approach on several computer vi sion problems.
****

The Multiverse Loss for Robust Transfer Learning
Etai Littwin, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3957-3966
Deep learning techniques are renowned for supporting effective transfer learning. However, as we demonstrate, the transferred representations support only a few modes of separation and much of its dimensionality is unutilized. In this work we suggest to learn, in the source domain, multiple orthogonal classifiers. We prove that this leads to a reduced rank representation, which however supports more discriminative directions. Interestingly, the softmax probabilities produced by the multiple classifiers are likely to be identical. Extensive experimental results further demonstrate the effectiveness of our method.
****

Learning From the Mistakes of Others: Matching Errors in Cross-Dataset Learning
Viktoriia Sharmanska, Novi Quadrianto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3967-3975
Can we learn about object classes in images by looking at a collection of relevant 3D models? Or if we want to learn about human (inter-)actions in images, can we benefit from videos or abstract illustrations that show these actions? A common aspect of these settings is the availability of additional or privileged data that can be exploited at training time and that will not be available and not of interest at test time. We seek to generalize the learning with privileged information (LUPI) framework, which requires additional information to be defined per image, to the setting where additional information is a data collection about the task of interest. Our framework minimises the distribution mismatch between errors made in images and in privileged data. The proposed method is tested on four publicly available datasets: Image+ClipArt, Image+3Dobject, and Image+Video. Experimental results reveal that our new LUPI paradigm naturally addresses the cross-dataset learning.
****

An Efficient Exact-PGA Algorithm for Constant Curvature Manifolds
Rudrasis Chakraborty, Dohyung Seo, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3976-3984
Manifold-valued datasets are widely encountered in many computer vision tasks. A non-linear analog of the PCA algorithm, called the Principal Geodesic Analysis (PGA) algorithm suited for data lying on Riemannian manifolds was reported in literature a decade ago. Since the objective function in the PGA algorithm is highly non-linear and hard to solve efficiently in general, researchers have proposed a linear approximation. Though this linear approximation is easy to compute, it lacks accuracy especially when the data exhibits a large variance. Recently, an alternative called the exact PGA was proposed which tries to solve the optimization without any linearization. For general Riemannian manifolds, though it yields a better accuracy than the original (linearized) PGA, for data that exhibit large variance, the optimization is not computationally efficient. In this paper, we propose an efficient exact PGA algorithm for constant curvature Riemannian manifolds (CCM-EPGA). The CCM-EPGA algorithm differs significantly from existing PGA algorithms in two aspects, (i) the distance between a given manifold-valued data point and the principal submanifold is computed analytically and thus no optimization is required as in the existing methods. (ii) Unlike the existing PGA algorithms, the descent into codimension-1 submanifolds does not require any optimization but is accomplished through the use of the Rimeannian inverse Exponential map and the parallel transport operations. We present theoretical and experimental results for constant curvature Riemannian manifolds depicting favorable performance of the CCM-EPGA algorithm compared to existing PGA algorithms. We also present data reconstruction from the principal components which has not been reported in literature in this setting.
****

Online Learning With Bayesian Classification Trees
Samuel Rota Bulo, Peter Kontschieder; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3985-3993
Randomized classification trees are among the most popular machine learning tool

s and found successful applications in many areas. Although this classifier was originally designed as offline learning algorithm, there has been an increased interest in the last years to provide an online variant. In this paper, we propose an online learning algorithm for classification trees that adheres to Bayesian principles. In contrast to state-of-the-art approaches that produce large forests with complex trees, we aim at constructing small ensembles consisting of shallow trees with high generalization capabilities. Experiments on benchmark machine learning and body part recognition datasets show superior performance over state-of-the-art approaches.

****

Cross-Stitch Networks for Multi-Task Learning

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3994-4003

Multi-task learning in Convolutional Networks has displayed remarkable success in the field of recognition. This success can be largely attributed to learning shared representations from multiple supervisory tasks. However, existing multi-task approaches rely on enumerating multiple network architectures specific to the tasks at hand, that do not generalize. In this paper, we propose a principled approach to learn shared representations in ConvNets using multi-task learning. Specifically, we propose a new sharing unit: "cross-stitch" unit. These units combine the activations from multiple networks and can be trained end-to-end. A network with cross-stitch units can learn an optimal combination of shared and task-specific representations. Our proposed method generalizes across multiple tasks and shows dramatically improved performance over baseline methods for categories with few training examples.

****

Deep Metric Learning via Lifted Structured Feature Embedding

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4004-4012

Learning the distance metric between pairs of examples is of great importance for learning and visual recognition. With the remarkable success from the state of the art convolutional neural networks, recent works have shown promising results on discriminatively training the networks to learn semantic feature embeddings where similar examples are mapped close to each other and dissimilar examples are mapped farther apart. In this paper, we describe an algorithm for taking full advantage of the training batches in the neural network training by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances. This step enables the algorithm to learn the state of the art feature embedding by optimizing a novel structured prediction objective for active hard negative mining on the lifted problem. Additionally, we collected Online Products data set: 120k images of 23k classes of online products for metric learning. Our experiments on the CUB-200-2011, CARS196, and Online Products datasets demonstrate significant improvement over existing deep feature embedding methods on all experimented embedding sizes with the GoogLeNet network. The source code and the data set are available at: https://github.com/rksltnl/Deep-Metric-Learning-CVPR16

****

Fast Algorithms for Convolutional Neural Networks

Andrew Lavin, Scott Gray; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4013-4021

Deep convolutional neural networks take GPU-days of computation to train on large data sets. Pedestrian detection for self driving cars requires very low latency. Image recognition for mobile phones is constrained by limited processing resources. The success of convolutional neural networks in these situations is limited by how fast we can compute them. Conventional FFT based convolution is fast for large filters, but state of the art convolutional neural networks use small, 3x3 filters. We introduce a new class of fast algorithms for convolutional neural networks using Winograd's minimal filtering algorithms. The algorithms compute minimal complexity convolution over small tiles, which makes them fast with sma

ll filters and small batch sizes. We benchmark a GPU implementation of our algor ithm with the VGG network and show state of the art throughput at batch sizes fr om 1 to 64.
****

Coordinating Multiple Disparity Proposals for Stereo Computation
Ang Li, Dapeng Chen, Yuanliu Liu, Zejian Yuan; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4022-4030
While great progress has been made in stereo computation over the last decades, large textureless regions remain challenging. Segment-based methods can tackle t his problem properly, but their performances are sensitive to the segmentation r esults. In this paper, we alleviate the sensitivity by generating multiple propo sals on absolute and relative disparities from multi-segmentations. These propos als supply rich descriptions of surface structures. Especially, the relative dis parity between distant pixels can encode the large structure, which is critical to handle the large texture-less regions. The proposals are coordinated by point -wise competition and pairwise collaboration within a MRF model. During inferenc e, a dynamic programming is performed in different directions with various step sizes, so the long-range connections are better preserved. In the experiments, w e carefully analyzed the effectiveness of the major components. Results on the 2 014 Middlebury and KITTI 2015 stereo benchmark show that our method is comparabl e to state-of-the-art.
****

Joint Multiview Segmentation and Localization of RGB-D Images Using Depth-Induce d Silhouette Consistency
Chi Zhang, Zhiwei Li, Rui Cai, Hongyang Chao, Yong Rui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4031-403 9
In this paper, we propose an RGB-D camera localization approach which takes an e ffective geometry constraint, i.e. silhouette consistency, into consideration. U nlike existing approaches which usually assume the silhouettes are provided, we consider more practical scenarios and generate the silhouettes for multiple view s on the fly. To obtain a set of accurate silhouettes, precise camera poses are required to propagate segmentation cues across views. To perform better localiza tion, accurate silhouettes are needed to constrain camera poses. Therefore the t wo problems are intertwined with each other and require a joint treatment. Facil itated by the available depth, we introduce a simple but effective silhouette co nsistency energy term that binds traditional appearance-based multiview segmenta tion cost and RGB-D frame-to-frame matching cost together. Optimization of the p roblem w.r.t. binary segmentation masks and camera poses naturally fits in the g raph cut minimization framework and the Gauss-Newton non-linear least-squares me thod respectively. Experiments show that the proposed approach achieves state-of -the-arts performance on both tasks of image segmentation and camera localizatio n.
****

A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation
Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexe y Dosovitskiy, Thomas Brox; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2016, pp. 4040-4048
Recent work has shown that optical flow estimation can be formulated as a superv ised learning task and can be successfully solved with convolutional networks. Training of the so-called FlowNet was enabled by a large synthetically generated dataset. The present paper extends the concept of optical flow estimation via c onvolutional networks to disparity and scene flow estimation. To this end, we p ropose three synthetic stereo video datasets with sufficient realism, variation, and size to successfully train large networks. Our datasets are the first larg e-scale datasets to enable training and evaluation of scene flow methods. Besid es the datasets, we present a convolutional network for real-time disparity esti mation that provides state-of-the-art results. By combining a flow and disparity estimation network and training it jointly, we demonstrate the first scene flow

estimation with a convolutional network.
****

6D Dynamic Camera Relocalization From Single Reference Image
Wei Feng, Fei-Peng Tian, Qian Zhang, Jizhou Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4049-4057
Dynamic relocalization of 6D camera pose from single reference image is a costly and challenging task that requires delicate hand-eye calibration and precision positioning platform to do 3D mechanical rotation and translation. In this paper, we show that high-quality camera relocalization can be achieved in a much less expensive way. Based on inexpensive platform with unreliable absolute repositioning accuracy (ARA), we propose a hand-eye calibration free strategy to actively relocate camera into the same 6D pose that produces the input reference image, by sequentially correcting 3D relative rotation and translation. We theoretically prove that, by this strategy, both rotational and translational relative pose can be effectively reduced to zero, with bounded unknown hand-eye pose displacement. To conquer 3D rotation and translation ambiguity, this theoretical strategy is further revised to a practical relocalization algorithm with faster convergence rate and more reliability by jointly adjusting 3D relative rotation and translation. Extensive experiments validate the effectiveness and superior accuracy of the proposed approach on laboratory tests and challenging real-world applications.
****

Dense Monocular Depth Estimation in Complex Dynamic Scenes
Rene Ranftl, Vibhav Vineet, Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4058-4066
We present an approach to dense depth estimation from a single monocular camera that is moving through a dynamic scene. The approach produces a dense depth map from two consecutive frames. Moving objects are reconstructed along with the surrounding environment. We provide a novel motion segmentation algorithm that segments the optical flow field into a set of motion models, each with its own epipolar geometry. We then show that the scene can be reconstructed based on these motion models by optimizing a convex program. The optimization jointly reasons about the scales of different objects and assembles the scene in a common coordinate frame, determined up to a global scale. Experimental results demonstrate that the presented approach outperforms prior methods for monocular depth estimation in dynamic scenes.
****

Using Self-Contradiction to Learn Confidence Measures in Stereo Vision
Christian Mostegel, Markus Rumpler, Friedrich Fraundorfer, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4067-4076
Learned confidence measures gain increasing importance for outlier removal and quality improvement in stereo vision. However, acquiring the necessary training data is typically a tedious and time consuming task that involves manual interaction, active sensing devices and/or synthetic scenes. To overcome this problem, we propose a new, flexible, and scalable way for generating training data that only requires a set of stereo images as input. The key idea of our approach is to use different view points for reasoning about contradictions and consistencies between multiple depth maps generated with the same stereo algorithm. This enables us to generate a huge amount of training data in a fully automated manner. Among other experiments, we demonstrate the potential of our approach by boosting the performance of three learned confidence measures on the KITTI2012 dataset by simply training them on a vast amount of automatically generated training data rather than a limited amount of laser ground truth data.
****

Understanding Real World Indoor Scenes With Synthetic Data
Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4077-4085

Scene understanding is a prerequisite to many high level tasks for any automated intelligent machine operating in real world environments. Recent attempts with supervised learning have shown promise in this direction but also highlighted the need for enormous quantity of supervised data --- performance increases in proportion to the amount of data used. However, this quickly becomes prohibitive when considering the manual labour needed to collect such data. In this work, we focus our attention on depth based semantic per-pixel labelling as a scene understanding problem and show the potential of computer graphics to generate virtually unlimited labelled data from synthetic 3D scenes. By carefully synthesizing training data with appropriate noise models we show comparable performance to state-of-the-art RGBD systems on NYUv2 dataset despite using only depth data as input and set a benchmark on depth-based segmentation on SUN RGB-D dataset.
****

Stereo Matching With Color and Monochrome Cameras in Low-Light Conditions
Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4086-4094
Consumer devices with stereo cameras have become popular because of their low-cost depth sensing capability. However, those systems usually suffer from low imaging quality and inaccurate depth acquisition under low-light conditions. To address the problem, we present a new stereo matching method with a color and monochrome camera pair. We focus on the fundamental trade-off that monochrome cameras have much better light-efficiency than color-filtered cameras. Our key ideas involve compensating for the radiometric difference between two cross-spectral images and taking full advantage of complementary data. Consequently, our method produces both an accurate depth map and high-quality images, which are applicable for various depth-aware image processing. Our method is evaluated using various datasets and the performance of our depth estimation consistently outperforms state-of-the-art methods.
****

Camera Calibration From Dynamic Silhouettes Using Motion Barcodes
Gil Ben-Artzi, Yoni Kasten, Shmuel Peleg, Michael Werman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4095-4103
Computing the epipolar geometry between cameras with very different viewpoints is often problematic as matching points are hard to find. In these cases, it has been proposed to use information from dynamic objects in the scene for suggesting point and line correspondences. We propose a speed up of about two orders of magnitude, as well as an increase in robustness and accuracy, to methods computing epipolar geometry from dynamic silhouettes based on a new temporal signature, motion barcode for lines. This is a binary temporal sequence for lines, indicating for each frame the existence of at least one foreground pixel on that line. The motion barcodes of two corresponding epipolar lines are very similar so the search for corresponding epipolar lines can be limited to lines having similar barcodes leading to increased speed, accuracy, and robustness in computing the epipolar geometry.
****

Structure-From-Motion Revisited
Johannes L. Schonberger, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104-4113
Incremental Structure-from-Motion is a prevalent strategy for 3D reconstruction from unordered image collections. While incremental reconstruction systems have tremendously advanced in all regards, robustness, accuracy, completeness, and scalability remain the key problems towards building a truly general-purpose pipeline. We propose a new SfM technique that improves upon the state of the art to make a further step towards this ultimate goal. The full reconstruction pipeline is released to the public as an open-source implementation.
****

Constructing Canonical Regions for Fast and Effective View Selection
Wencheng Wang, Tianhao Gao; Proceedings of the IEEE Conference on Computer Visio

n and Pattern Recognition (CVPR), 2016, pp. 4114-4122
In view selection, little work has been done for optimizing the search process; views must be densely distributed and checked individually. Thus, evaluating poo r views wastes much time, and a poor view may even be misidentified as a best on e. In this paper, we propose a search strategy by identifying the regions that a re very likely to contain best views, referred to as canonical regions. It is by decomposing the model under investigation into meaningful parts, and using the canonical views of these parts to generate canonical regions. Applying existing view selection methods in the canonical regions can not only accelerate the sear ch process but also guarantee the quality of obtained views. As a result, when o ur canonical regions are used for searching N-best views during comprehensive mo del analysis, we can attain greater search speed and reduce the number of views required. Experimental results show the effectiveness of our method.
****

Prior-Less Compressible Structure From Motion
Chen Kong, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2016, pp. 4123-4131
Many non-rigid 3D structures are not modelled well through a low-rank subspace a ssumption. This is problematic when it comes to their reconstruction through St ructure from Motion (SfM). We argue in this paper that a more expressive and ge neral assumption can be made around compressible 3D structures. The vision comm unity, however, has hitherto struggled to formulate effective strategies for rec overing such structures after projection without the aid of additional priors (e .g. temporal ordering, rigid substructures, etc.). In this paper we present a " prior-less" approach to solve compressible SfM. Specifically, we demonstrate ho w the problem of SfM - assuming compressible 3D structures - can be theoreticall y characterized as a block sparse dictionary learning problem. We validate our approach experimentally by demonstrating reconstructions of 3D structures that a re intractable using current state-of-the-art low-rank SfM approaches.
****

Rolling Shutter Camera Relative Pose: Generalized Epipolar Geometry
Yuchao Dai, Hongdong Li, Laurent Kneip; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2016, pp. 4132-4140
The vast majority of modern consumer-grade cameras employ a rolling shutter mech anism. In dynamic geometric computer vision applications such as visual SLAM, th e so-called rolling shutter effect therefore needs to be properly taken into acc ount. A dedicated relative pose solver appears to be the first problem to solve, as it is of eminent importance to bootstrap any derivation of multi-view geomet ry. However, despite its significance, it has received inadequate attention to d ate. This paper presents a detailed investigation of the geometry of the rollin g shutter relative pose problem. We introduce the rolling shutter essential matr ix, and establish its link to existing models such as the push-broom cameras, su mmarized in a clean hierarchy of multi-perspective cameras. The generalization o f well-established concepts from epipolar geometry is completed by a definition of the Sampson distance in the rolling shutter case. The work is concluded with a careful investigation of the introduced epipolar geometry for rolling shutter cameras on several dedicated benchmarks.
****

Structure From Motion With Objects
Marco Crocco, Cosimo Rubino, Alessio Del Bue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4141-4149
This paper shows for the first time that is possible to reconstruct the position of rigid objects and to jointly recover affine camera calibration solely from a set of object detections in a video sequence. In practice, this work can be con sidered as the extension of Tomasi and Kanade factorization method using objects . Instead of using points to form a rank constrained measurement matrix, we can form a matrix with similar rank properties using 2D object detection proposals. In detail, we first fit an ellipse onto the image plane at each bounding box as given by the object detector. The collection of all the ellipses in the dual spa ce is used to create a measurement matrix that gives a specific rank constraint.

This matrix can be factorised and metrically upgraded in order to provide the affine camera matrices and the 3D position of the objects as an ellipsoid. Moreover, we recover the full 3D quadric thus giving additional information about object occupancy and 3D pose. Finally, we also show that 2D points measurements can be seamlessly included in the framework to reduce the number of objects required. This last aspect unifies the classical point-based Tomasi and Kanade approach with objects in a unique framework. Experiments with synthetic and real data show the feasibility of our approach for the affine camera case.
****

DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed With Deep Features

Ayan Sinha, Chiho Choi, Karthik Ramani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4150-4158

We propose DeepHand to estimate the 3D pose of a hand using depth data from commercial 3D sensors. We discriminatively train convolutional neural networks to output a low dimensional activation feature given a depth map. This activation feature vector is representative of the global or local joint angle parameters of a hand pose. We efficiently identify 'spatial' nearest neighbors to the activation feature, from a database of features corresponding to synthetic depth maps, and store some 'temporal' neighbors from previous frames. Our matrix completion algorithm uses these 'spatio-temporal' activation features and the corresponding known pose parameter values to to estimate the unknown pose parameters of the input feature vector. Our database of activation features supplements large viewpoint coverage and our hierarchical estimation of pose parameters is robust to occlusions. We show that our approach compares favorably to state-of-the-art methods while achieving real time performance (32 FPS) on a standard computer.
****

Multi-Oriented Text Detection With Fully Convolutional Networks

Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, Xiang Bai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4159-4167

In this paper, we propose an unconventional approach for text detection in natural images. Both global and local cues are taken into account for localizing text lines in a coarse-to-fine procedure. First, a Fully Convolutional Network (FCN) model is trained for predicting a salient map of text regions in a holistic manner. Then, a set of hypotheses text lines are estimated by combining the salient map and MSER components. Finally, another FCN classifier is used for predicting the centroid of each character, in order to remove the false hypotheses. The framework is general for handling texts in multiple orientations, languages and fonts. The proposed method consistently achieves the state-of-the-art performance on three text detection benchmarks: MSRA-TD500, ICDAR2015, and ICDAR2013.
****

Robust Scene Text Recognition With Automatic Rectification

Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4168-4176

Recognizing text in natural images is a challenging task with many unsolved problems. Different from those in documents, words in natural images often possess irregular shapes, which are caused by perspective distortion, curved character placement, etc. We propose RARE (Robust text recognizer with Automatic REctification), a recognition model that is robust to irregular text. RARE is a specially-designed deep neural network, which consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN). In testing, an image is firstly rectified via a predicted Thin-Plate-Spline (TPS) transformation, into a more "readable" image for the following SRN, which recognizes text through a sequence recognition approach. We show that the model is able to recognize several types of irregular text, including perspective text and curved text. RARE is end-to-end trainable, requiring only images and associated text labels, making it convenient to train and deploy the model in practical systems. State-of-the-art or highly-competitive performance achieved on several benchmarks well demonstrates the effe

ctiveness of the proposed model.
****

Mnemonic Descent Method: A Recurrent Process Applied for End-To-End Face Alignme nt

George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, S tefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pat tern Recognition (CVPR), 2016, pp. 4177-4187

Cascaded regression has recently become the method of choice for solving non-lin ear least squares problems such as deformable image alignment. Given a sizeable training set, cascaded regression learns a set of generic rules that are sequent ially applied to minimise the least squares problem. Despite the success of cas caded regression for problems such as face alignment and head pose estimation, t here are several shortcomings arising in the strategies proposed thus far. Spec ifically, (a) the regressors are learnt independently, (b) the descent direction s may cancel one another out and (c) handcrafted features (e.g., HoGs, SIFT etc .) are mainly used to drive the cascade, which may be sub-optimal for the task a t hand. In this paper, we propose a combined and jointly trained convolutional r ecurrent neural network architecture that allows the training of an end-to-end t o system that attempts to alleviate the aforementioned drawbacks. The recurrent module facilitates the joint optimisation of the regressors by assuming the casc ades form a nonlinear dynamical system, in effect fully utilising the informatio n between all cascade levels by introducing a memory unit that shares informatio n across all levels. The convolutional module allows the network to extract feat ures that are specialised for the task at hand and are experimentally shown to o utperform hand-crafted features. We show that the application of the proposed ar chitecture for the problem of face alignment results in a strong improvement ove r the current state-of-the-art.

****

Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting

Amin Jourabloo, Xiaoming Liu; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2016, pp. 4188-4196

Large-pose face alignment is a very challenging problem in computer vision, whic h is used as a prerequisite for many important vision tasks, e.g, face recogniti on and 3D face reconstruction. Recently, there have been a few attempts to solve this problem, but still more research is needed to achieve highly accurate resu lts. In this paper, we propose a face alignment method for large-pose face image s, by combining the powerful cascaded CNN regressor method and 3DMM. We formulat e the face alignment as a 3DMM fitting problem, where the camera projection matr ix and 3D shape parameters are estimated by a cascade of CNN-based regressors. T he dense 3D shape allows us to design pose-invariant appearance features for eff ective CNN learning. Extensive experiments are conducted on the challenging data bases (AFLW and AFW), with comparison to the state of the art.

****

Adaptive 3D Face Reconstruction From Unconstrained Photo Collections

Joseph Roth, Yiying Tong, Xiaoming Liu; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2016, pp. 4197-4206

Given a collection of "in-the-wild" face images captured under a variety of unkn own pose, expression, and illumination conditions, this paper presents a method for reconstructing a 3D face surface model of an individual along with albedo in formation. Motivated by the success of recent face reconstruction techniques on large photo collections, we extend prior work to adapt to low quality photo coll ections with fewer images. We achieve this by fitting a 3D Morphable Model to fo rm a personalized template and developing a novel photometric stereo formulation , under a coarse-to-fine scheme. Superior experimental results are reported on s ynthetic and real-world photo collections.

****

Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D C onvolutional Neural Network

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, Jan Ka utz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogniti

on (CVPR), 2016, pp. 4207-4215
Automatic detection and classification of dynamic hand gestures in real-world sy stems intended for human computer interaction is challenging as: 1) there is a l arge diversity in how people perform gestures, making detection and classificati on difficult; 2) the system must work online in order to avoid noticeable lag be tween performing a gesture and its classification; in fact, a negative lag (clas sification before the gesture is finished) is desirable, as feedback to the user can then be truly instantaneous. In this paper, we address these challenges wit h a recurrent three-dimensional convolutional neural network that performs simul taneous detection and classification of dynamic hand gestures from multi-modal d ata. We employ connectionist temporal classification to train the network to pre dict class labels from in-progress gestures in unsegmented input streams. In ord er to validate our method, we introduce a new challenging multi-modal dynamic ha nd gesture dataset captured with depth, color and stereo-IR sensors. On this cha llenging dataset, our gesture recognition system achieves an accuracy of 83.8%, outperforms competing state-of-the-art algorithms, and approaches human accuracy of 88.4%. Moreover, our method achieves state-of-the-art performance on SKIG an d ChaLearn2014 benchmarks.
****
Kinematic Structure Correspondences via Hypergraph Matching
Hyung Jin Chang, Tobias Fischer, Maxime Petit, Martina Zambelli, Yiannis Demiris ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4216-4225
In this paper, we present a novel framework for finding the kinematic structure correspondence between two objects in videos via hypergraph matching. In contras t to prior appearance and graph alignment based matching methods which have been applied among two similar static images, the proposed method finds corresponden ces between two dynamic kinematic structures of heterogeneous objects in videos. Our main contributions can be summarised as follows: (i) casting the kinematic structure correspondence problem into a hypergraph matching problem, incorporati ng multi-order similarities with normalising weights, (ii) a structural topology similarity measure by a new topology constrained subgraph isomorphism aggregati on, (iii) a kinematic correlation measure between pairwise nodes, and (iv) a com binatorial local motion similarity measure using geodesic distance on the Rieman nian manifold. We demonstrate the robustness and accuracy of our method through a number of experiments on complex articulated synthetic and real data.
****
CP-mtML: Coupled Projection Multi-Task Metric Learning for Large Scale Face Retr ieval
Binod Bhattarai, Gaurav Sharma, Frederic Jurie; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4226-4235
We propose a novel Coupled Projection multi-task Met- ric Learning (CP-mtML) met hod for large scale face re- trieval. In contrast to previous works which were l imited to low dimensional features and small datasets, the proposed method scale s to large datasets with high dimensional face descriptors. It utilises pairwise (dis-)similarity constraints as supervision and hence does not require exhausti ve class annotation for every training image. While, traditionally, multi-task l earning methods have been validated on same dataset but different tasks, we work on the more chal- lenging setting with heterogeneous datasets and different tas ks. We show empirical validation on multiple face im- age datasets of different facial traits, e.g. identity, age and expression. We use classic Local Binary Pa ttern (LBP) de- scriptors along with the recent Deep Convolutional Neural Networ k (CNN) features. The experiments clearly demon- strate the scalability and impr oved performance of the pro- posed method on the tasks of identity and age based face image retrieval compared to competitive existing methods, on the standard datasets and with the presence of a million distractor face images.
****
PatchBatch: A Batch Augmented Loss for Optical Flow
David Gadot, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2016, pp. 4236-4245

We propose a new pipeline for optical flow computation, based on Deep Learning t
echniques. We suggest using a Siamese CNN to independently, and in parallel, com
pute the descriptors of both images. The learned descriptors are then compared e
fficiently using the L2 norm and do not require network processing of patch pair
s. The success of the method is based on an innovative loss function that comput
es higher moments of the loss distributions for each training batch. Combined wi
th an Approximate Nearest Neighbor patch matching method and a flow interpolatio
n technique, state of the art performance is obtained on the most challenging an
d competitive optical flow benchmarks.
****

Joint Recovery of Dense Correspondence and Cosegmentation in Two Images
Tatsunori Taniai, Sudipta N. Sinha, Yoichi Sato; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4246-4255
We propose a new technique to jointly recover cosegmentation and dense per-pixel
 correspondence in two images. Our method parameterizes the correspondence field
 using piecewise similarity transformations and recovers a mapping between the e
stimated common "foreground" regions in the two images allowing them to be preci
sely aligned. Our formulation is based on a hierarchical Markov random field mod
el with segmentation and transformation labels. The hierarchical structure uses
nested image regions to constrain inference across multiple scales. Unlike prior
 hierarchical methods which assume that the structure is given, our proposed ite
rative technique dynamically recovers the structure as a variable along with the
 labeling. This joint inference is performed in an energy minimization framework
 using iterated graph cuts. We evaluate our method on a new dataset of 400 image
 pairs with manually obtained ground truth, where it outperforms state-of-the-ar
t methods designed specifically for either cosegmentation or correspondence esti
mation.
****

Multi-View People Tracking via Hierarchical Trajectory Composition
Yuanlu Xu, Xiaobai Liu, Yang Liu, Song-Chun Zhu; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4256-4265
This paper presents a hierarchical composition approach for multi-view object tr
acking. The key idea is to adaptively exploit multiple cues in both 2D and 3D, e
.g., ground occupancy consistency, appearance similarity, motion coherence etc.,
 which are mutually complementary while tracking the humans of interests over ti
me. While feature online selection has been extensively studied in the past lite
rature, it remains unclear how to effectively schedule these cues for the tracki
ng purpose especially when encountering various challenges, e.g. occlusions, con
junctions, and appearance variations. To do so, we propose a hierarchical compos
ition model and re-formulate multi-view multi-object tracking as a problem of co
mpositional structure optimization. We setup a set of composition criteria, each
 of which corresponds to one particular cue. The hierarchical composition proces
s is pursued by exploiting different criteria, which impose constraints between
a graph node and its offsprings in the hierarchy. We learn the composition crite
ria using MLE on annotated data and efficiently construct the hierarchical graph
 by an iterative greedy pursuit algorithm. In the experiments, we demonstrate su
perior performance of our approach on three public datasets, one of which is new
ly created by us to test various challenges in multi-view multi-object tracking.
****

Object Tracking via Dual Linear Structured SVM and Explicit Feature Map
Jifeng Ning, Jimei Yang, Shaojie Jiang, Lei Zhang, Ming-Hsuan Yang; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,
pp. 4266-4274
Structured support vector machine (SSVM) based methods has demonstrated encourag
ing performance in recent object tracking benchmarks. However, the complex and e
xpensive optimization limits their deployment in real-world applications. In thi
s paper, we present a simple yet efficient dual linear SSVM (DLSSVM) algorithm t
o enable fast learning and execution during tracking. By analyzing the dual vari
ables, we propose a primal classifier update formula where the learning step siz
e is computed in closed form. This online learning method significantly improves

the robustness of the proposed linear SSVM with low computational cost. Second, we approximate the intersection kernel for feature representations with an explicit feature map to further improve tracking performance. Finally, we extend the proposed DLSSVM tracker in a multiscale manner to address the "drift" problem. Experimental results on large benchmark datasets with 50 and 100 video sequences show that the proposed DLSSVM tracking algorithm achieves state-of-the-art performance.

****

Robust, Real-Time 3D Tracking of Multiple Objects With Similar Appearances
Taiki Sekii; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4275-4283

This paper proposes a novel method for tracking multiple moving objects and recovering their three-dimensional (3D) models separately using multiple calibrated cameras. For robustly tracking objects with similar appearances, the proposed method uses geometric information regarding 3D scene structure rather than appearance. A major limitation of previous techniques is foreground confusion, in which the shapes of objects and/or ghosting artifacts are ignored and are hence not appropriately specified in foreground regions. To overcome this limitation, our method classifies foreground voxels into targets (objects and artifacts) in each frame using a novel, probabilistic two-stage framework. This is accomplished by step-wise application of a track graph describing how targets interact and the maximum a posteriori expectation-maximization algorithm for the estimation of target parameters. We introduce mixture models with semiparametric component distributions regarding 3D target shapes. In order to not confuse artifacts with objects of interest, we automatically detect and track artifacts based on a closed-world assumption. Experimental results show that our method outperforms state-of-the-art trackers on seven public sequences while achieving real-time performance.

****

An Egocentric Look at Video Photographer Identity
Yedid Hoshen, Shmuel Peleg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4284-4292

Egocentric cameras are being worn by an increasing number of users, among them many security forces worldwide. GoPro cameras already penetrated the mass market, reporting substantial increase in sales every year. As head-worn cameras do not capture the photographer, it may seem that the anonymity of the photographer is preserved even when the video is publicly distributed. We show that camera motion, as can be computed from the egocentric video, provides unique identity information. The photographer can be reliably recognized from a few seconds of video captured when walking. The proposed method achieves more than 90% recognition accuracy in cases where the random success rate is only 3%. Applications can include theft prevention by locking the camera when not worn by its lawful owner. Searching video sharing services (e.g. YouTube) for egocentric videos shot by a specific photographer may also become possible. An important message in this paper is that photographers should be aware that sharing egocentric video will compromise their anonymity, even when their face is not visible.

****

Learning Multi-Domain Convolutional Neural Networks for Visual Tracking
Hyeonseob Nam, Bohyung Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4293-4302

We propose a novel visual tracking algorithm based on the representations from a discriminatively trained Convolutional Neural Network (CNN). Our algorithm pretrains a CNN using a large set of videos with tracking ground-truths to obtain a generic target representation. Our network is composed of shared layers and multiple branches of domain-specific layers, where domains correspond to individual training sequences and each branch is responsible for binary classification to identify target in each domain. We train each domain in the network iteratively to obtain generic target representations in the shared layers. When tracking a target in a new sequence, we construct a new network by combining the shared layers in the pretrained CNN with a new binary classification layer, which is updated online. Online tracking is performed by evaluating the candidate windows random

ly sampled around the previous target state. The proposed algorithm illustrates outstanding performance in existing tracking benchmarks.
****

Hedged Deep Tracking
Yuankai Qi, Shengping Zhang, Lei Qin, Hongxun Yao, Qingming Huang, Jongwoo Lim, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4303-4311
In recent years, several methods have been developed to utilize hierarchical features learned from a deep convolutional neural network (CNN) for visual tracking. However, as the features from a certain CNN layer characterize an object of interest from only one aspect or one level, the performance of such trackers trained with features from one layer (usually the last second layer) can be further improved. In this paper, we propose a novel CNN based tracking framework, which takes full advantage of features from different CNN layers and uses an adaptive Hedge method to hedge several CNN trackers into a stronger one. Extensive experiments on a benchmark dataset of 100 challenging image sequences demonstrate the effectiveness of the proposed algorithm compared with several state-of-the-art trackers.
****

Structural Correlation Filter for Robust Visual Tracking
Si Liu, Tianzhu Zhang, Xiaochun Cao, Changsheng Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4312-4320
In this paper, we propose a novel structural correlation filter (SCF) model for robust visual tracking. The proposed SCF model takes part-based tracking strategies into account in a correlation filter tracker, and exploits circular shifts of all parts for their motion modeling to preserve target object structure. Compared with existing correlation filter trackers, our proposed tracker has several advantages: (1) Due to the part strategy, the learned structural correlation filters are less sensitive to partial occlusion, and have computational efficiency and robustness. (2) The learned filters are able to not only distinguish the parts from the background as the traditional correlation filters, but also exploit the intrinsic relationship among local parts via spatial constraints to preserve object structure. (3) The learned correlation filters not only make most parts share similar motion, but also tolerate outlier parts that have different motion. Both qualitative and quantitative evaluations on challenging benchmark image sequences demonstrate that the proposed SCF tracking algorithm performs favorably against several state-of-the-art methods.
****

Visual Tracking Using Attention-Modulated Disintegration and Integration
Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, Jin Young Choi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4321-4330
In this paper, we present a novel attention-modulated visual tracking algorithm that decomposes an object into multiple cognitive units, and trains multiple elementary trackers in order to modulate the distribution of attention according to various feature and kernel types. In the integration stage it recombines the units to memorize and recognize the target object effectively. With respect to the elementary trackers, we present a novel attentional feature-based correlation filter (AtCF) that focuses on distinctive attentional features. The effectiveness of the proposed algorithm is validated through experimental comparison with state-of-the-art methods on widely-used tracking benchmark datasets.
****

A Continuous Occlusion Model for Road Scene Understanding
Vikas Dhiman, Quoc-Huy Tran, Jason J. Corso, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4331-4339
We present a physically interpretable, continuous 3D model for handling occlusions with applications to road scene understanding. We probabilistically assign each point in space to an object with a theoretical modeling of the reflection and transmission probabilities for the corresponding camera ray. Our modeling is un

ified in handling occlusions across a variety of scenarios, such as associating structure from motion point tracks with potentially occluded objects or modeling object detection scores in applications such as 3D localization. For point track association, our model uniformly handles static and dynamic objects, which is an advantage over motion segmentation approaches traditionally used in multibody SFM. Detailed experiments on the KITTI dataset show the superiority of the proposed method over both state-of-the-art motion segmentation and a baseline that heuristically uses detection bounding boxes for resolving occlusions. We also demonstrate how our continuous occlusion model may be applied to the task of 3D localization in road scenes.
****

Virtual Worlds as Proxy for Multi-Object Tracking Analysis
Adrien Gaidon, Qiao Wang, Yohann Cabon, Eleonora Vig; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4340-4349
Modern computer vision algorithms typically require expensive data acquisition and accurate manual labeling. In this work, we instead leverage the recent progress in computer graphics to generate fully labeled, dynamic, and photo-realistic proxy virtual worlds. We propose an efficient real-to-virtual world cloning method, and validate our approach by building and publicly releasing a new video dataset, called Virtual KITTI, automatically labeled with accurate ground truth for object detection, tracking, scene and instance segmentation, depth, and optical flow. We provide quantitative experimental evidence suggesting that (i) modern deep learning algorithms pre-trained on real data behave similarly in real and virtual worlds, and (ii) pre-training on virtual data improves performance. As the gap between real and virtual worlds is small, virtual worlds enable measuring the impact of various weather and imaging conditions on recognition performance, all other things being equal. We show these factors may affect drastically otherwise high-performing deep models for tracking.
****

Uncalibrated Photometric Stereo by Stepwise Optimization Using Principal Components of Isotropic BRDFs
Keisuke Midorikawa, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4350-4358
The uncalibrated photometric stereo problem for non-Lambertian surfaces is challenging because of the large number of unknowns and its ill-posed nature stemming from unknown reflectance functions. We propose a model that represents various isotropic reflectance functions by using the principal components of items in a dataset, and formulate the uncalibrated photometric stereo as a regression problem. We then solve it by stepwise optimization utilizing principal components in order of their importance. We have also developed two techniques that lead to convergence and highly accurate reconstruction, namely (1) a coarse-to-fine approach with normal grouping, and (2) a randomized multipoint search. Our experimental results with synthetic data showed that our method significantly outperformed previous methods. We also evaluated the algorithm in terms of real image data, where it gave good reconstruction results.
****

Unbiased Photometric Stereo for Colored Surfaces: A Variational Approach
Yvain Queau, Roberto Mecca, Jean-Denis Durou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4359-4368
3D shape recovery using photometric stereo (PS) gained increasing attention in the computer vision community in the last three decades due to its ability to recover the thinnest geometric structures. Yet, the reliabiliy of PS for color images is difficult to guarantee, because existing methods are usually formulated as the sequential estimation of the colored albedos, the normals and the depth. Hence, the overall reliability depends on that of each subtask. In this work we propose a new formulation of color photometric stereo, based on image ratios, that makes the technique independent from the albedos. This allows the unbiased 3D-reconstruction of colored surfaces in a single step, by solving a system of linear PDEs using a variational approach.

****
3D Reconstruction of Transparent Objects With Position-Normal Consistency
Yiming Qian, Minglun Gong, Yee Hong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4369-4377
Estimating the shape of transparent and refractive objects is one of the few open problems in 3D reconstruction. Under the assumption that the rays refract only twice when traveling through the object, we present the first approach to simultaneously reconstructing the 3D positions and normals of the object's surface at both refraction locations. Our acquisition setup requires only two cameras and one monitor, which serves as the light source. After acquiring the ray-ray correspondences between each camera and the monitor, we solve an optimization function which enforces a new position-normal consistency constraint. That is, the 3D positions of surface points shall agree with the normals required to refract the rays under Snell's law. Experimental results using both synthetic and real data demonstrate the robustness and accuracy of the proposed approach.
****
Real-Time Depth Refinement for Specular Objects
Roy Or-El, Rom Hershkovitz, Aaron Wetzler, Guy Rosman, Alfred M. Bruckstein, Ron Kimmel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4378-4386
The introduction of consumer RGB-D scanners set off a major boost in 3D computer vision research. Yet, the precision of existing depth scanners is not accurate enough to recover fine details of a scanned object. While modern shading based depth refinement methods have been proven to work well with Lambertian objects, they break down in the presence of specularities. We present a novel shape from shading framework that addresses this issue and enhances both diffuse and specular objects' depth profiles. We take advantage of the built-in monochromatic IR projector and IR images of the RGB-D scanners and present a lighting model that accounts for the specular regions in the input image. Using this model, we reconstruct the depth map in real-time. Both quantitative tests and visual evaluations prove that the proposed method produces state of the art depth reconstruction results.
****
Recovering Transparent Shape From Time-Of-Flight Distortion
Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4387-4395
This paper presents a method for recovering shape and normal of a transparent object from a single viewpoint using a Time-of-Flight (ToF) camera. Our method is built upon the fact that the speed of light varies with the refractive index of the medium and therefore the depth measurement of a transparent object with a ToF camera may be distorted. We show that, from this ToF distortion, the refractive light path can be uniquely determined by estimating a single parameter. We estimate this parameter by introducing a surface normal consistency between the one determined by a light path candidate and the other computed from the corresponding shape. The proposed method is evaluated by both simulation and real-world experiments and shows faithful transparent shape recovery.
****
Robust Light Field Depth Estimation for Noisy Scene With Occlusion
W. Williem, In Kyu Park; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4396-4404
Light field depth estimation is an essential part of many light field applications. Numerous algorithms have been developed using various light field characteristics. However, conventional methods fail when handling noisy scene with occlusion. To remedy this problem, we present a light field depth estimation method which is more robust to occlusion and less sensitive to noise. Novel data costs using angular entropy metric and adaptive defocus response are introduced. Integration of both data costs improves the occlusion and noise invariant capability significantly. Cost volume filtering and graph cut optimization are utilized to improve the accuracy of the depth map. Experimental results confirm that the propos

ed method is robust and achieves high quality depth maps in various scenes. The proposed method outperforms the state-of-the-art light field depth estimation methods in qualitative and quantitative evaluation.
****

Rotational Crossed-Slit Light Field
Nianyi Li, Haiting Lin, Bilin Sun, Mingyuan Zhou, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4405-4413
Light fields (LFs) are image-based representation that records the radiance along all rays along every direction through every point in space. Traditionally LFs are acquired by using a 2D grid of evenly spaced pinhole cameras or by translating a pinhole camera along the 2D grid using a robot arm. In this paper, we present a novel LF sampling scheme by exploiting a special non-centric camera called the crossed-slit or XSlit camera. An XSlit camera acquires rays that simultaneously pass through two oblique slits. We show that, instead of translating the camera as in the pinhole case, we can effectively sample the LF by rotating individual or both slits while keeping the camera fixed. This leads a "fixed-location" LF acquisition scheme. We further show through theoretical analysis and experiments that the resulting XSlit LFs provide several advantages: they provide more dense spatial-angular sampling, are amenable multi-view stereo matching and volumetric reconstruction, and can synthesize unique refocusing effects.
****

Single Image Object Modeling Based on BRDF and R-Surfaces Learning
Fabrizio Natola, Valsamis Ntouskos, Fiora Pirri, Marta Sanzari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4414-4423
A methodology for 3D surface modeling from a single image is proposed. The principal novelty is concave and specular surface modeling without any externally imposed prior. The main idea of the method is to use BRDFs and generated rendered surfaces, to transfer the normal field, computed for the generated samples, to the unknown surface. The transferred information is adequate to blow and sculpt the segmented image mask in to a bas-relief of the object. The object surface is further refined basing on a photo-consistency formulation that relates for error minimization the original image and the modeled object.
****

A Nonlinear Regression Technique for Manifold Valued Data With Applications to Medical Image Analysis
Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, Michael S. Okun, David E. Viallancourt, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4424-4432
Regression is an essential tool in Statistical analysis of data with many applications in Computer Vision, Machine Learning, Medical Imaging and various disciplines of Science and Engineering. Linear and nonlinear regression in a vector space setting has been well studied in literature. However, generalizations to manifold-valued data are only recently gaining popularity. With the exception of a few, most existing methods of regression for manifold valued data are limited to geodesic regression which is a generalization of the linear regression in vector-spaces. In this paper, we present a novel nonlinear kernel-based regression method that is applicable to manifold valued data. Our method is applicable to cases when the independent and dependent variables in the regression model are both manifold-valued or one is manifold-valued and the other is vector or scalar valued. Further, unlike most methods, our method does not require any imposed ordering on the manifold-valued data. The performance of our model is tested on a large number of real data sets acquired from Alzhiemers and movement disorder (Parkinsons and Essential Tremor) patients. We present an extensive set of results along with statistical validation and comparisons.
****

RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian With Application to Material Recognition
Qilong Wang, Peihua Li, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE Confere

nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4433-4441
Infinite dimensional covariance descriptors can provide richer and more discrimi
native information than their low dimensional counterparts. In this paper, we pr
opose a novel image descriptor, namely, robust approximate infinite dimensional
Gaussian (RAID-G). The challenges of RAID-G mainly lie on two aspects: (1) descr
iption of infinite dimensional Gaussian is difficult due to its non-linear Riema
nnian geometric structure and the infinite dimensional setting, hence effective
approximation is necessary; (2) traditional maximum likelihood estimation (MLE)
is not robust to high (even infinite) dimensional covariance matrix in Gaussian
setting. To address these challenges, explicit feature mapping (EFM) is first in
troduced for effective approximation of infinite dimensional Gaussian induced by
 additive kernel function, and then a new regularized MLE method based on von Ne
umann divergence is proposed for robust estimation of covariance matrix. The EFM
 and proposed regularized MLE allow a closed-form of RAID-G, which is very effic
ient and effective for high dimensional features. We extend RAID-G by using the
outputs of deep convolutional neural networks as original features, and apply it
 to material recognition. Our approach is evaluated on five material benchmarks
and one fine-grained benchmark. It achieves 84.9% accuracy on FMD and 86.3% accu
racy on UIUC material database, which are much higher than state-of-the-arts.
****

An Empirical Evaluation of Current Convolutional Architectures' Ability to Manag
e Nuisance Location and Scale Variability
Nikolaos Karianakis, Jingming Dong, Stefano Soatto; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4442-4451
We conduct an empirical study to test the ability of convolutional neural networ
ks (CNNs) to reduce the effects of nuisance transformations of the input data, s
uch as location, scale and aspect ratio. We isolate factors by adopting a common
 convolutional architecture either deployed globally on the image to compute cla
ss posterior distributions, or restricted locally to compute class conditional d
istributions given location, scale and aspect ratios of bounding boxes determine
d by proposal heuristics. In theory, averaging the latter should yield inferior
performance compared to proper marginalization. Yet empirical evidence suggests
the converse, leading us to conclude that - at the current level of complexity o
f convolutional architectures and scale of the data sets used to train them - CN
Ns are not very effective at marginalizing nuisance variability. We also quantif
y the effects of context on the overall classification task and its impact on th
e performance of CNNs, and propose improved sampling techniques for heuristic pr
oposal schemes that improve end-to-end performance to state-of-the-art levels. W
e test our hypothesis on a classification task using the ImageNet Challenge benc
hmark and on a wide-baseline matching task using the Oxford and Fischer's datase
ts.
****

Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilate
ral Neural Networks
Varun Jampani, Martin Kiefel, Peter V. Gehler; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4452-4461
Bilateral filters have wide spread use due to their edge-preserving properties.
The common use case is to manually choose a parametric filter type, usually a Ga
ussian filter. In this paper, we will generalize the parametrization and in part
icular derive a gradient descent algorithm so the filter parameters can be learn
ed from data. This derivation allows to learn high dimensional linear filters th
at operate in sparsely populated feature spaces. We build on the permutohedral l
attice construction for efficient filtering. The ability to learn more general f
orms of high-dimensional filters can be used in several diverse applications. Fi
rst, we demonstrate the use in applications where single filter applications are
 desired for runtime reasons. Further, we show how this algorithm can be used to
 learn the pairwise potentials in densely connected conditional random fields an
d apply these to different image segmentation tasks. Finally, we introduce layer
s of bilateral filters in CNN and propose bilateral neural networks for the use
of high-dimensional sparse data. This view provides new ways to encode model str

ucture into network architectures. A diverse set of experiments empirically validates the usage of general forms of filters.
****

Mixture of Bilateral-Projection Two-Dimensional Probabilistic Principal Component Analysis
Fujiao Ju, Yanfeng Sun, Junbin Gao, Simeng Liu, Yongli Hu, Baocai Yin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4462-4470
The probabilistic principal component analysis (PPCA) is built upon a global linear mapping, with which it is insufficient to model complex data variation. This paper proposes a mixture of bilateral-projection probabilistic principal component analysis model (mixB2DPPCA) on 2D data. With multi-components in the mixture, this model can be seen as a `soft' cluster algorithm and has capability of modeling data with complex structures. A Bayesian inference scheme has been proposed based on the variational EM (Expectation-Maximization) approach for learning model parameters. Experiments on some publicly available databases show that the performance of mixB2DPPCA has been largely improved, resulting in more accurate reconstruction errors and recognition rates than the existing PCA-based algorithms.
****

Rolling Rotations for Recognizing Human Actions From 3D Skeletal Data
Raviteja Vemulapalli, Rama Chellapa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4471-4479
Recently, skeleton-based human action recognition has been receiving significant attention from various research communities due to the availability of depth sensors and real-time depth-based 3D skeleton estimation algorithms. In this work, we use rolling maps for recognizing human actions from 3D skeletal data. The rolling map is a well-defined mathematical concept that has not been explored much by the vision community. First, we represent each skeleton using the relative 3D rotations between various body parts. Since 3D rotations are members of the special orthogonal group SO(3), our skeletal representation becomes a point in the Lie group SO(3) X ... X SO(3), which is also a Riemannian manifold. Then, using this representation, we model human actions as curves in this Lie group. Since classification of curves in this non-Euclidean space is a difficult task, we unwrap the action curves onto the Lie algebra (which is a vector space) by combining the logarithm map with rolling maps, and perform classification in the Lie algebra. Experimental results on three action datasets show that the proposed approach performs equally well or better when compared to state-of-the-art.
****

Improving the Robustness of Deep Neural Networks via Stability Training
Stephan Zheng, Yang Song, Thomas Leung, Ian Goodfellow; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4480-4488
In this paper we address the issue of output instability of deep neural networks: small perturbations in the visual input can significantly distort the feature embeddings and output of a neural network. Such instability affects many deep architectures with state-of-the-art performance on a wide range of computer vision tasks. We present a general stability training method to stabilize deep networks against small input distortions that result from various types of common image processing, such as compression, rescaling, and cropping. We validate our method by stabilizing the state-of-the-art Inception architecture against these types of distortions. In addition, we demonstrate that our stabilized model gives robust state-of-the-art performance on large-scale near-duplicate detection, similar-image ranking, and classification on noisy datasets.
****

Logistic Boosting Regression for Label Distribution Learning
Chao Xing, Xin Geng, Hui Xue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4489-4497
Label Distribution Learning (LDL) is a general learning framework which includes both single label and multi-label learning as its special cases. One of the mai

n assumptions made in traditional LDL algorithms is the derivation of the parame tric model as the maximum entropy model. While it is a reasonable assumption wit hout additional information, there is no particular evidence supporting it in th e problem of LDL. Alternatively, using a general LDL model family to approximate this parametric model can avoid the potential influence of the specific model. In order to learn this general model family, this paper uses a method called Log istic Boosting Regression (LogitBoost) which can be seen as an additive weighted function regression from the statistical viewpoint. For each step, we can fit i ndividual weighted regression function (base learner) to realize the optimizatio n gradually. The base learners are chosen as weighted regression tree and vector tree, which constitute two algorithms named LDLogitBoost and AOSO-LDLogitBoost in this paper. Experiments on facial expression recognition, crowd opinion predi ction on movies and apparent age estimation show that LDLogitBoost and AOSO-LDLo gitBoost can achieve better performance than traditional LDL algorithms as well as other LogitBoost algorithms.

****

Efficient Temporal Sequence Comparison and Classification Using Gram Matrix Embe ddings on a Riemannian Manifold

Xikang Zhang, Yin Wang, Mengran Gou, Mario Sznaier, Octavia Camps; Proceedings o f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p p. 4498-4507

In this paper we propose a new framework to compare and classify temporal sequen ces. The proposed approach captures the underlying dynamics of the data while av oiding expensive estimation procedures, making it suitable to process large numb ers of sequences. The main idea is to first embed the sequences into a Riemann ian manifold by using positive definite regularized Gram matrices of their Han kelets. The advantages of the this approach are: 1) it allows for using non-Eu clidean similarity functions on the Positive Definite matrix manifold, which ca pture better the underlying geometry than directly comparing the sequences or th eir Hankel matrices; and 2) Gram matrices inherit desirable properties from th e underlying Hankel matrices: their rank measure the complexity of the underlyi ng dynamics, and the rank and the coefficients of the associated regressive mode ls are invariant to affine transformations and varying initial conditions. The benefits of this approach are illustrated with extensive experiments in 3D actio n recognition using 3D joints sequences. In spite of its simplicity, the perfo rmance of this approach is competitive or better than using state-of-art appr oaches for this problem. Further, these results hold across a variety of metri cs, supporting the idea that the improvement stems from the embedding itself, ra ther than from using one of these metrics.

****

Deep Reflectance Maps

Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, Tinne Tuy telaars; Proceedings of the IEEE Conference on Computer Vision and Pattern Recog nition (CVPR), 2016, pp. 4508-4516

Undoing the image formation process and therefore decomposing appearance into it s intrinsic properties is a challenging task due to the under-constraint nature of this inverse problem. While significant progress has been made on inferring s hape, materials and illumination from images only, progress in unconstrained set ting is still limited. We propose a fully convolutional neural architecture to e stimate reflectance maps of specular materials in natural lighting conditions. W e achieve this in an end-to-end learning formulation that directly predicts a re flectance map from the image itself. We show how to improve estimates by facilit ating additional supervision in an indirect scheme that first predicts surface o rientation and afterwards predicts the reflectance map by a learning-based spars e data interpolation. In order to analyze performance on this difficult task, we propose a new challenge of Specular MAterials on SHapes with complex IllumiNati on (SMASHINg) using both synthetic and real images. Furthermore, we show the app lication our method to a range of image-based editing tasks on real images.

****

Semantic Filtering

Qingxiong Yang; Proceedings of the IEEE Conference on Computer Vision and Patter n Recognition (CVPR), 2016, pp. 4517-4526
Edge-preserving image operations aim at smoothing an image without blurring the edges. Many excellent edge-preserving filtering techniques have been proposed re cently to reduce the computational complexity or/and separate different scale st ructures. They normally adopt a user-selected scale measurement to control the d etail/texture smoothing. However, natural photos contain objects of different si zes which cannot be described by a single scale measurement. On the other hand, edge/contour detection/analysis is closely related to edge-preserving filtering and has achieved significant progress recently. Nevertheless, most of the state-of-the-art filtering techniques ignore the success in this area. Inspired by the fact that learning-based edge detectors/classifiers significantly outperform tr aditional manually-designed detectors, this paper proposes a learning-based edge -preserving filtering technique. It synergistically combines the efficiency of t he recursive filter and the effectiveness of the recent edge detector for scale-aware edge-preserving filtering. Unlike previous filtering methods, the propose filter can efficiently extract subjectively-meaningful structures from natural s cenes containing multiple-scale objects.
****

UAV Sensor Fusion With Latent-Dynamic Conditional Random Fields in Coronal Plane Estimation
Amir M. Rahimi, Raphael Ruschel, B.S. Manjunath; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4527-4534
We present a real-time body orientation estimation in a micro-Unmanned Air Vehic le video stream. This work is part of a fully autonomous UAV system which can ma neuver to face a single individual in challenging outdoor environments. Our body orientation estimation consists of the following steps: (a) obtaining a set of visual appearance models for each body orientation, where each model is tagged w ith a set of scene information (obtained from sensors); (b) exploiting the mutua l information of on-board sensors using latent-dynamic conditional random fields (LDCRF); (c) Characterizing each visual appearance model with the most discrimi native sensor information; (d) fast estimation of body orientation during the te st flights given the LDCRF parameters and the corresponding sensor readings. The key aspects of our approach is to add sparsity to the sensor readings with late nt variables followed by long range dependency analysis. Experimental results ob tained over real-time video streams demonstrate a significant improvement in bot h speed (15-fps) and accuracy (72%) compared to the state of the art techniques that only rely on visual data. Video demonstration of our autonomous flights (bo th from ground view and aerial view) are included in the supplementary material.
****

Robust Visual Place Recognition With Graph Kernels
Elena Stumm, Christopher Mei, Simon Lacroix, Juan Nieto, Marco Hutter, Roland Si egwart; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogn ition (CVPR), 2016, pp. 4535-4544
A novel method for visual place recognition is introduced and evaluated, demonst rating robustness to perceptual aliasing and observation noise. This is achieved by increasing discrimination through a more structured representation of visual observations. Estimation of observation likelihoods are based on graph kernel f ormulations, utilizing both the structural and visual information encoded in cov isibility graphs. The proposed probabilistic model is able to circumvent the typ ically difficult and expensive posterior normalization procedure by exploiting t he information available in visual observations. Furthermore, the place recognit ion complexity is independent of the size of the map. Results show improvements over the state-of-the-art on a diverse set of both public datasets and novel exp eriments, highlighting the benefit of the approach.
****

Semantic Image Segmentation With Task-Specific Edge Detection Using CNNs and a D iscriminatively Trained Domain Transform
Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, Alan L. Y uille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni

tion (CVPR), 2016, pp. 4545-4554
Deep convolutional neural networks (CNNs) are the backbone of state-of-art seman tic image segmentation systems. Recent work has shown that complementing CNNs wi th fully-connected conditional random fields (CRFs) can significantly enhance th eir object localization accuracy, yet dense CRF inference is computationally exp ensive. We propose replacing the fully-connected CRF with domain transform (DT), a modern edge-preserving filtering method in which the amount of smoothing is c ontrolled by a reference edge map. Domain transform filtering is several times f aster than dense CRF inference and we show that it yields comparable semantic se gmentation results, accurately capturing object boundaries. Importantly, our for mulation allows learning the reference edge map from intermediate CNN features i nstead of using the image gradient magnitude as in standard DT filtering. This p roduces task-specific edges in an end-to-end trainable system optimizing the tar get semantic segmentation quality.
****

Natural Language Object Retrieval
Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, Trevor Darrel l; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4555-4564
In this paper, we address the task of natural language object retrieval, to loca lize a target object within a given image based on a natural language query of t he object. Natural language object retrieval differs from text-based image retri eval task as it involves spatial information about objects within the scene and global scene context. To address this issue, we propose a novel Spatial Context Recurrent ConvNet (SCRC) model as scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information into the network. Our model processes query text, local image descr iptors, spatial configurations and global context features through a recurrent n etwork, outputs the probability of the query text conditioned on each candidate box as a score for the box, and can transfer visual-linguistic knowledge from im age captioning domain to our task. Experimental results demonstrate that our met hod effectively utilizes both local and global information, outperforming previo us baseline methods significantly on different datasets and scenarios, and can e xploit large scale vision and language datasets for knowledge transfer.
****

DenseCap: Fully Convolutional Localization Networks for Dense Captioning
Justin Johnson, Andrej Karpathy, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4565-4574
We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consis t of a single word, and Image Captioning when one predicted region covers the fu ll image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an i mage with a single, efficient forward pass, requires no external regions proposa ls, and can be trained end-to-end with a single round of optimization. The archi tecture is composed of a Convolutional Network, a novel dense localization layer , and Recurrent Neural Network language model that generates the label sequences . We evaluate our network on the Visual Genome dataset, which comprises 94,000 i mages and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in bot h generation and retrieval settings.
****

Unsupervised Learning From Narrated Instruction Videos
Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Lapt ev, Simon Lacoste-Julien; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4575-4583
We address the problem of automatically learning the main steps to complete a ce rtain task, such as changing a car tire, from a set of narrated instruction vide os. The contributions of this paper are three-fold. First, we develop a new unsu

pervised learning approach that takes advantage of the complementary nature of t
he input video and the associated narration. The method solves two clustering pr
oblems, one in text and one in video, applied one after each other and linked by
 joint constraints to obtain a single coherent sequence of steps in both modalit
ies. Second, we collect and annotate a new challenging dataset of real-world ins
truction videos from the Internet. The dataset contains about 800,000 frames for
 five different tasks that include complex interactions between people and objec
ts, and are captured in a variety of indoor and outdoor settings. Third, we expe
rimentally demonstrate that the proposed method can automatically discover, in a
n unsupervised manner, the main steps to achieve the task and locate the steps i
n the input videos.
****

Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks
Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, Wei Xu; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4584-4593
We present an approach that exploits hierarchical Recurrent Neural Networks (RNN
s) to tackle the video captioning problem, i.e., generating one or multiple sent
ences to describe a realistic video. Our hierarchical framework contains a sente
nce generator and a paragraph generator. The sentence generator produces one sim
ple short sentence that describes a specific short video interval. It exploits b
oth temporal- and spatial-attention mechanisms to selectively focus on visual el
ements during generation. The paragraph generator captures the inter-sentence de
pendency by taking as input the sentential embedding produced by the sentence ge
nerator, combining it with the paragraph history, and outputting the new initial
 state for the sentence generator. We evaluate our approach on two large-scale b
enchmark datasets: YouTubeClips and TACoS-MultiLevel. The experiments demonstrat
e that our approach significantly outperforms the current state-of-the-art metho
ds with BLEU@4 scores 0.499 and 0.305 respectively.
****

Jointly Modeling Embedding and Translation to Bridge Video and Language
Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, Yong Rui; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4594-4602
Automatically describing video content with natural language is a fundamental ch
allenge of computer vision. Recurrent Neural Networks (RNNs), which models seque
nce dynamics, has attracted increasing attention on visual interpretation. Howev
er, most existing approaches generate a word locally with the given previous wor
ds and the visual content, while the relationship between sentence semantics and
 visual content is not holistically exploited. As a result, the generated senten
ces may be contextually correct but the semantics (e.g., subjects, verbs or obje
cts) are not true. This paper presents a novel unified framework, named Long Sh
ort-Term Memory with visual-semantic Embedding (LSTM-E), which can simultaneousl
y explore the learning of LSTM and visual-semantic embedding. The former aims to
 locally maximize the probability of generating the next word given previous wor
ds and visual content, while the latter is to create a visual-semantic embedding
 space for enforcing the relationship between the semantics of the entire senten
ce and visual content. The experiments on YouTube2Text dataset show that our pro
posed LSTM-E achieves to-date the best published performance in generating natur
al sentences: 45.3% and 31.0% in terms of BLEU@4 and METEOR, respectively. Super
ior performances are also reported on two movie description datasets (M-VAD and
MPII-MD). In addition, we demonstrate that LSTM-E outperforms several state-of-t
he-art techniques in predicting Subject-Verb-Object (SVO) triplets.
****

We Are Humor Beings: Understanding and Predicting Visual Humor
Arjun Chandrasekaran, Ashwin K. Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhru
v Batra, C. Lawrence Zitnick, Devi Parikh; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4603-4612
Humor is an integral part of human lives. Despite being tremendously impactful,
it is perhaps surprising that we do not have a detailed understanding of humor y
et. As interactions between humans and AI systems increase, it is imperative tha
t these systems are taught to understand subtleties of human expressions such as

humor. In this work, we are interested in the question - what content in a scene causes it to be funny? As a first step towards understanding visual humor, we analyze the humor manifested in abstract scenes and design computational models for them. We collect two datasets of abstract scenes that facilitate the study of humor at both the scene-level and the object-level. We analyze the funny scenes and explore the different types of humor depicted in them via human studies. We model two tasks that we believe demonstrate an understanding of some aspects of visual humor. The tasks involve predicting the funniness of a scene and altering the funniness of a scene. We show that our models perform well quantitatively, and qualitatively through human studies. Our datasets are publicly available.
****

Where to Look: Focus Regions for Visual Question Answering
Kevin J. Shih, Saurabh Singh, Derek Hoiem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4613-4621
We present a method that learns to answer visual questions by selecting image regions relevant to the text-based query. Our method maps textual queries and visual features from various regions into a shared space where they are compared for relevance with an inner product. Our method exhibits significant improvements in answering questions such as "what color," where it is necessary to evaluate a specific location, and "what room," where it selectively identifies informative image regions. Our model is tested on the recently released VQA dataset, which features free-form human-annotated questions and answers.
****

Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge From External Sources
Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4622-4630
We propose a method for visual question answering which combines an internal representation of the content of an image with information extracted from a general knowledge base to answer a broad range of image-based questions. This allows more complex questions to be answered using the predominant neural network-based approach than has previously been possible. It particularly allows questions to be asked about the contents of an image, even when the image itself does not contain the whole answer. The method constructs a textual representation of the semantic content of an image, and merges it with textual information sourced from a knowledge base, to develop a deeper understanding of the scene viewed. Priming a recurrent neural network with this combined information, and the submitted question, leads to a very flexible visual question answering approach. We are specifically able to answer questions posed in natural language, that refer to information not contained in the image. We demonstrate the effectiveness of our model on two publicly available datasets, Toronto COCO-QA and VQA, and show that it produces the best reported results in both cases.
****

MovieQA: Understanding Stories in Movies Through Question-Answering
Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4631-4640
We introduce the MovieQA dataset which aims to evaluate automatic story comprehension from both video and text. The dataset consists of 14,944 questions about 408 movies with high semantic diversity. The questions range from simpler "Who" did "What" to "Whom", to "Why" and "How" certain events occurred. Each question comes with a set of five possible answers; a correct one and four deceiving answers provided by human annotators. Our dataset is unique in that it contains multiple sources of information -- video clips, plots, subtitles, scripts, and DVS. We analyze our data through various statistics and methods. We further extend existing QA techniques to show that question-answering with such open-ended semantics is hard. We make this data set public along with an evaluation benchmark to encourage inspiring work in this challenging domain.
****

TGIF: A New Dataset and Benchmark on Animated GIF Description
Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4641-4650
With the recent popularity of animated GIFs on social media, there is need for ways to index them with rich metadata. To advance research on animated GIF understanding, we collected a new dataset, Tumblr GIF (TGIF), with 100K animated GIFs from Tumblr and 120K natural language descriptions obtained via crowdsourcing. The motivation for this work is to develop a testbed for image sequence description systems, where the task is to generate natural language descriptions for animated GIFs or video clips. To ensure a high quality dataset, we developed a series of novel quality controls to validate free-form text input from crowdworkers. We show that there is unambiguous association between visual content and natural language descriptions in our dataset, making it an ideal benchmark for the visual content captioning task. We perform extensive statistical analyses to compare our dataset to existing image and video description datasets. Next, we provide baseline results on the animated GIF description task, using three representative techniques: nearest neighbor, statistical machine translation, and recurrent neural networks. Finally, we show that models fine-tuned from our animated GIF description dataset can be helpful for automatic movie description.
****

Image Captioning With Semantic Attention
Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651-4659
Automatically generating a natural language description of an image has attracted interests recently both because of its importance in practical applications and because it connects two major artificial intelligence fields: computer vision and natural language processing. Existing approaches are either top-down, which start from a gist of an image and convert it into words, or bottom-up, which come up with words describing various aspects of an image and then combine them. In this paper, we propose a new algorithm that combines both approaches through a model of semantic attention. Our algorithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. We evaluate our algorithm on two public benchmarks: Microsoft COCO and Flickr30K. Experimental results show that our algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics.
****

Temporally Coherent 4D Reconstruction of Complex Dynamic Scenes
Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, Adrian Hilton; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4660-4669
This paper presents an approach for reconstruction of 4D temporally coherent models of complex dynamic scenes. No prior knowledge is required of scene structure or camera calibration allowing reconstruction from multiple moving cameras. Sparse-to-dense temporal correspondence is integrated with joint multi-view segmentation and reconstruction to obtain a complete 4D representation of static and dynamic objects. Temporal coherence is exploited to overcome visual ambiguities resulting in improved reconstruction of complex scenes. Robust joint segmentation and reconstruction of dynamic objects is achieved by introducing a geodesic star convexity constraint. Comparative evaluation is performed on a variety of unstructured indoor and outdoor dynamic scenes with hand-held cameras and multiple people. This demonstrates reconstruction of complete temporally coherent 4D scene models with improved non-rigid object segmentation and shape reconstruction.
****

Consensus of Non-Rigid Reconstructions
Minsik Lee, Jungchan Cho, Songhwai Oh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4670-4678

Recently, there have been many progresses for the problem of non-rigid structure reconstruction based on 2D trajectories, but it is still challenging to deal with complex deformations or restricted view ranges. Promising alternatives are the piecewise reconstruction approaches, which divide trajectories into several local parts and stitch their individual reconstructions to produce an entire 3D structure. These methods show the state-of-the-art performance, however, most of them are specialized for relatively smooth surfaces and some are quite complicated. Meanwhile, it has been reported numerously in the field of pattern recognition that obtaining consensus from many weak hypotheses can give a strong, powerful result. Inspired by these reports, in this paper, we push the concept of part-based reconstruction to the limit: Instead of considering the parts as explicitly-divided local patches, we draw a large number of small random trajectory sets. From their individual reconstructions, we pull out a statistic of each 3D point to retrieve a strong reconstruction, of which the procedure can be expressed as a sparse l_1-norm minimization problem. In order to resolve the reflection ambiguity between weak (and possibly bad) reconstructions, we propose a novel optimization framework which only involves a single eigenvalue decomposition. The proposed method can be applied to any type of data and outperforms the existing methods for the benchmark sequences, even though it is composed of a few, simple steps. Furthermore, it is easily parallelizable, which is another advantage.
****

Isometric Non-Rigid Shape-From-Motion in Linear Time
Shaifali Parashar, Daniel Pizarro, Adrien Bartoli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4679-4687
We study Isometric Non-Rigid Shape-from-Motion (Iso-NRSfM): given multiple intrinsically calibrated monocular images, we want to reconstruct the time-varying 3D shape of an object undergoing isometric deformations. We show that Iso-NRSfM is solvable from the warps (the inter-image geometric transformations). We propose a new theoretical framework based on Riemmanian manifolds to represent the unknown 3D surfaces, as embeddings of the camera's retinal planes. This allows us to use the manifolds' metric tensor and Christoffel Symbol fields, which we prove are related across images by simple rules depending only on the warps. This forms a set of important theoretical results. Using the infinitesimal planarity formulation, it then allows us to derive a system of two quartics in two variables for each image pair. The sum-of-squares of these polynomials is independent of the number of images and can be solved globally, forming a well-posed problem for N >= 3 images, whose solution directly leads to the surface's normal field. The proposed method outperforms existing work in terms of accuracy and computation cost on synthetic and real datasets.
****

Learning Online Smooth Predictors for Realtime Camera Planning Using Recurrent Decision Trees
Jianhui Chen, Hoang M. Le, Peter Carr, Yisong Yue, James J. Little; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4688-4696
We study the problem of online prediction for realtime camera planning, where the goal is to predict smooth trajectories that correctly track and frame objects of interest (e.g., players in a basketball game). The conventional approach for training predictors does not directly consider temporal consistency, and often produces undesirable jitter. Although post-hoc smoothing (e.g., via a Kalman filter) can mitigate this issue to some degree, it is not ideal due to overly stringent modeling assumptions (e.g., Gaussian noise). We propose a recurrent decision tree framework that can directly incorporate temporal consistency into a data-driven predictor, as well as a learning algorithm that can efficiently learn such temporally smooth models. Our approach does not require any post-processing, making online smooth predictions much easier to generate when the noise model is unknown. We apply our approach to sports broadcasting: given noisy player detections, we learn where the camera should look based on human demonstrations. Our experiments exhibit significant improvements over conventional baselines and showcase the practicality of our approach.

****
Egocentric Future Localization
Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, Jianbo Shi; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4697-4705
We presents a method for future localization: to predict plausible future trajec tories of ego-motion in egocentric stereo images. Our paths avoid obstacles, mo ve between objects, even turn around a corner into space behind objects. As a by product of the predicted trajectories, we discover the empty space occluded by f oreground objects. One key innovation is the creation of an EgoRetinal map, ak in to an illustrated tourist map, that `rearranges' pixels taking into accounts depth information, the ground plane, and body motion direction, so that it allow s motion planning and perception of objects on one image space. We learn to plan trajectories directly on this EgoRetinal map using first person experience of walking around in a variety of scenes. In a testing phase, given an novel scene , we find multiple hypotheses of future trajectories from the learned experience . We refine them by minimizing a cost function that describes compatibility bet ween the obstacles in the EgoRetinal map and trajectories. We quantitatively ev aluate our method to show predictive validity and apply to various real world da ily activities including walking, shopping, and social interactions.
****
Full Flow: Optical Flow Estimation By Global Optimization Over Regular Grids
Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2016, pp. 4706-4714
We present a global optimization approach to optical flow estimation. The approa ch optimizes a classical optical flow objective over the full space of mappings between discrete grids. No descriptor matching is used. The highly regular struc ture of the space of mappings enables optimizations that reduce the computationa l complexity of the algorithm's inner loop from quadratic to linear and support efficient matching of tens of thousands of nodes to tens of thousands of displac ements. We show that one-shot global optimization of a classical Horn-Schunck-ty pe objective over regular grids at a single resolution is sufficient to initiali ze continuous interpolation and achieve state-of-the-art performance on challeng ing modern benchmarks.
****
Structured Feature Learning for Pose Estimation
Xiao Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4715-4723
In this paper, we propose a structured feature learning framework to reason the correlation among body joints at the feature level in human pose estimation. Dif ferent from existing approaches of modeling structures on score maps or predicte d labels, feature maps preserve substantially richer descriptions of body joints . The relationships between feature maps of joints are captured with the introdu ced geometrical transform kernels, which can be easily implemented with a convol ution layer. Features and their relationships are jointly learned in an end-to-e nd learning system. A bi-directional tree structured model is proposed, so that the feature channels at a body joint can well receive information from other joi nts. The proposed framework improves feature learning substantially. With very s imple post processing, it reaches the best mean PCP on the LSP and FLIC datasets . Compared with the baseline of learning features at each joint separately with ConvNet, the mean PCP has been improved by 18% on FLIC. The code is released to the public.
****
Convolutional Pose Machines
Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4724 -4732
Pose Machines provide a sequential prediction framework for learning rich implic it spatial models. In this work we show a systematic design for how convolutiona l networks can be incorporated into the pose machine framework for learning imag e features and image-dependent spatial models for the task of pose estimation. T

he contribution of this paper is to implicitly model long-range dependencies bet
ween variables in structured prediction tasks such as articulated pose estimatio
n. We achieve this by designing a sequential architecture composed of convolutio
nal networks that directly operate on belief maps from previous stages, producin
g increasingly refined estimates for part locations, without the need for explic
it graphical model-style inference. Our approach addresses the characteristic di
fficulty of vanishing gradients during training by providing a natural learning
objective function that enforces intermediate supervision, thereby replenishing
back-propagated gradients and conditioning the learning procedure. We demonstrat
e state-of-the-art performance and outperform competing methods on standard benc
hmarks including the MPII, LSP, and FLIC datasets.
****

Human Pose Estimation With Iterative Error Feedback
Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, Jitendra Malik; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,
pp. 4733-4742
Hierarchical feature extractors such as Convolutional Networks (ConvNets) have a
chieved impressive performance on a variety of classification tasks using purely
feedforward processing. Feedforward architectures can learn rich representation
s of the input space but do not explicitly model dependencies in the output spac
es, that are quite structured for tasks such as articulated human pose estimatio
n or object segmentation. Here we propose a framework that expands the expressiv
e power of hierarchical feature extractors to encompass both input and output sp
aces, by introducing top-down feedback. Instead of directly predicting the outpu
ts in one go, we use a self-correcting model that progressively changes an initi
al solution by feeding back error predictions, in a process we call Iterative Er
ror Feedback (IEF). IEF shows excellent performance on the task of articulated p
ose estimation in the challenging MPII and LSP benchmarks, matching the state-of
-the-art without requiring ground truth scale annotation.
****

WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks
Thibaut Durand, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4743-4752
In this paper, we introduce a novel framework for WEakly supervised Learning of
Deep cOnvolutional neural Networks (WELDON). Our method is dedicated to automati
cally selecting relevant image regions from weak annotations, e.g. global image
labels, and encompasses the following contributions. Firstly, WELDON leverages r
ecent improvements on the Multiple Instance Learning paradigm, i.e. negative ev
idence scoring and top instance selection. Secondly, the deep CNN is trained to
optimize Average Precision, and fine-tuned on the target dataset with efficient
computations due to convolutional feature sharing. A thorough experimental valid
ation shows that WELDON outperforms state-of-the-art results on six different da
tasets.
****

DisturbLabel: Regularizing CNN on the Loss Layer
Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, Qi Tian; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4753-47
62
During a long period of time we are combating over-fitting in the CNN training p
rocess with model regularization, including weight decay, model averaging, data
augmentation, etc. In this paper, we present DisturbLabel, an extremely simple a
lgorithm which randomly replaces a part of labels as incorrect values in each it
eration. Although it seems weird to intentionally generate incorrect training la
bels, we show that DisturbLabel prevents the network training from over-fitting
by implicitly averaging over exponentially many networks which are trained with
different label sets. To the best of our knowledge, DisturbLabel serves as the f
irst work which adds noises on the loss layer. Meanwhile, DisturbLabel cooperate
s well with Dropout to provide complementary regularization functions. Experimen
ts demonstrate competitive recognition results on several popular image recognit
ion datasets.

****
Gradual DropIn of Layers to Train Very Deep Neural Networks
Leslie N. Smith, Emily M. Hand, Timothy Doster; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4763-4771
We introduce the concept of dynamically growing a neural network during training . In particular, an untrainable deep network starts as a trainable shallow netwo rk and newly added layers are slowly, organically added during training, thereby increasing the network's depth. This is accomplished by a new layer, which we c all DropIn. The DropIn layer starts by passing the output from a previous layer (effectively skipping over the newly added layers), then increasingly including units from the new layers for both feedforward and backpropagation. We show that deep networks, which are untrainable with conventional methods, will converge w ith DropIn layers interspersed in the architecture. In addition, we demonstrate that DropIn provides regularization during training in an analogous way as dropo ut. Experiments are described with the MNIST dataset and various expanded LeNet architectures, CIFAR-10 dataset with its architecture expanded from 3 to 11 laye rs, and on the ImageNet dataset with the AlexNet architecture expanded to 13 lay ers and the VGG 16-layer architecture.
****
Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition
Zhiwei Deng, Arash Vahdat, Hexiang Hu, Greg Mori; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4772-4781
Rich semantic relations are important in a variety of visual recognition problem s. As a concrete example, group activity recognition involves the interactions and relative spatial relations of a set of people in a scene. State of the art r ecognition methods center on deep learning approaches for training highly effect ive, complex classifiers for interpreting images. However, bridging the relativ ely low-level concepts output by these methods to interpret higher-level composi tional scenes remains a challenge. Graphical models are a standard tool for thi s task. In this paper, we propose a method to integrate graphical models and dee p neural networks into a joint framework. Instead of using a traditional inferen ce method, we use a sequential inference modeled by a recurrent neural network. Beyond this, the appropriate structure for inference can be learned by imposing gates on edges between nodes. Empirical results on group activity recognition d emonstrate the potential of this model to handle highly structured learning task s.
****
Deep SimNets
Nadav Cohen, Or Sharir, Amnon Shashua; Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition (CVPR), 2016, pp. 4782-4791
We present a deep layered architecture that generalizes convolutional neural net works (ConvNets). The architecture, called SimNets, is driven by two operators: (i) a similarity function that generalizes inner-product, and (ii) a log-mean-e xp function called MEX that generalizes maximum and average. The two operators applied in succession give rise to a standard neuron but in "feature space". Th e feature spaces realized by SimNets depend on the choice of the similarity oper ator. The simplest setting, which corresponds to a convolution, realizes the fe ature space of the Exponential kernel, while other settings realize feature spac es of more powerful kernels (Generalized Gaussian, which includes as special cas es RBF and Laplacian), or even dynamically learned feature spaces (Generalized M ultiple Kernel Learning). As a result, the SimNet contains a higher abstraction level compared to a traditional ConvNet. We argue that enhanced expressiveness is important when the networks are small due to run-time constraints (such as t hose imposed by mobile applications). Empirical evaluation validates the superi or expressiveness of SimNets, showing a significant gain in accuracy over ConvNe ts when computational resources at run-time are limited. We also show that in l arge-scale settings, where computational complexity is less of a concern, the ad ditional capacity of SimNets can be controlled with proper regularization, yield ing accuracies comparable to state of the art ConvNets.

****
Studying Very Low Resolution Recognition Using Deep Networks
Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4792-4800
Visual recognition research often assumes a sufficient resolution of the region of interest (ROI). That is usually violated in practice, inspiring us to explore the Very Low Resolution Recognition (VLRR) problem. Typically, the ROI in a VLRR problem can be smaller than 16 x16 pixels, and is challenging to be recognized even by human experts. We attempt to solve the VLRR problem using deep learning methods. Taking advantage of techniques primarily in super resolution, domain adaptation and robust regression, we formulate a dedicated deep learning method and demonstrate how these techniques are incorporated step by step. Any extra complexity, when introduced, is fully justified by both analysis and simulation results. The resulting Robust Partially Coupled Networks achieves feature enhancement and recognition simultaneously. It allows for both the flexibility to combat the LR-HR domain mismatch, and the robustness to outliers. Finally, the effectiveness of the proposed models is evaluated on three different VLRR tasks, including face identification, digit recognition and font recognition, all of which obtain very impressive performances.
****
Deep Gaussian Conditional Random Field Network: A Model-Based Deep Network for Discriminative Denoising
Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4801-4809
We propose a novel end-to-end trainable deep network architecture for image denoising based on a Gaussian Conditional Random Field (GCRF) model. In contrast to the existing discriminative denoising methods that train a separate model for each individual noise level, the proposed deep network explicitly models the input noise variance and hence is capable of handling a range of noise levels. Our deep network, which we refer to as deep GCRF network, consists of two sub-networks: (i) a parameter generation network that generates the pairwise potential parameters based on the noisy input image, and ii) an inference network whose layers perform the computations involved in an iterative GCRF inference procedure. We train two deep GCRF networks (each network operates over a range of noise levels: one for low input noise levels and one for high input noise levels) discriminatively by maximizing the peak signal-to-noise ratio measure. Experiments on Berkeley segmentation and PASCALVOC datasets show that the proposed approach produces results on par with the state-of-the-art without training a separate network for each individual noise level.
****
Event-Specific Image Importance
Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, Garrison W. Cottrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4810-4819
When creating a photo album of an event, people typically select a few important images to keep or share. There is some consistency in the process of choosing the important images, and discarding the unimportant ones. Modeling this selection process will assist automatic photo selection and album summarization. In this paper, we show that the selection of important images is consistent among different viewers, and that this selection process is related to the event type of the album. We introduce the concept of event-specific image importance. We collected a new event album dataset with human annotation of the relative image importance with each event album. We also propose a Convolutional Neural Network (CNN) based method to predict the image importance score of a given event album, using a novel rank loss function and a progressive training scheme. Results demonstrate that our method significantly outperforms various baseline methods.
****
Quantized Convolutional Neural Networks for Mobile Devices
Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, Jian Cheng; Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4820-4828
Recently, convolutional neural networks (CNN) have demonstrated impressive performance in various computer vision tasks. However, high performance hardware is typically indispensable for the application of CNN models due to the high computation complexity, which prohibits their further extensions. In this paper, we propose an efficient framework, namely Quantized CNN, to simultaneously speed-up the computation and reduce the storage and memory overhead of CNN models. Both filter kernels in convolutional layers and weighting matrices in fully-connected layers are quantized, aiming at minimizing the estimation error of each layer's response. Extensive experiments on the ILSVRC-12 benchmark demonstrate 4 6x speed-up and 15 20x compression with merely one percentage loss of classification accuracy. With our quantized CNN model, even mobile devices can accurately classify images within one second.
****

Inverting Visual Representations With Convolutional Networks
Alexey Dosovitskiy, Thomas Brox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4829-4837
Feature representations, both hand-designed and learned ones, are often hard to analyze and interpret, even when they are extracted from visual data. We propose a new approach to study image representations by inverting them with an up-convolutional neural network. We apply the method to shallow representations (HOG, SIFT, LBP), as well as to deep networks. For shallow representations our approach provides significantly better reconstructions than existing methods, revealing that there is surprisingly rich information contained in these features. Inverting a deep network trained on ImageNet provides several insights into the properties of the feature representation learned by the network. Most strikingly, the colors and the rough contours of an image can be reconstructed from activations in higher network layers and even from the predicted class probabilities.
****

Pose-Aware Face Recognition in the Wild
Iacopo Masi, Stephen Rawls, Gerard Medioni, Prem Natarajan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4838-4846
We propose a method to push the frontiers of unconstrained face recognition in the wild, focusing on the problem of extreme pose variations. As opposed to current techniques which either expect a single model to learn pose invariance through massive amounts of training data, or which normalize images to a single frontal pose, our method explicitly tackles pose variation by using multiple pose-specific models and rendered face images. We leverage deep Convolutional Neural Networks (CNNs) to learn discriminative representations we call Pose-Aware Models (PAMs) using 500K images from the CASIA WebFace dataset. We present a comparative evaluation on the new IARPA Janus Benchmark A (IJB-A) and PIPA datasets. On these datasets PAMs achieve remarkably better performance than commercial products and surprisingly also outperform methods that are specifically fine-tuned on the target dataset.
****

Multi-View Deep Network for Cross-View Classification
Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4847-4855
Cross-view recognition that intends to classify samples between different views is an important problem in computer vision. The large discrepancy between different even heterogenous views make this problem quite challenging. To eliminate the complex (maybe even highly nonlinear) view discrepancy for favorable cross-view recognition, we propose a multi-view deep network (MvDN), which seeks for a non-linear discriminant and view-invariant representation shared between multiple views. Specifically, our proposed MvDN network consists of two sub-networks, view-specific sub-network attempting to remove view-specific variations and the following common sub-network attempting to obtain common representation shared by all views. As the objective of MvDN network, the Fisher loss, i.e. the Rayleigh q

uotient objective, is calculated from the samples of all views so as to guide th e learning of the whole network. As a result, the representation from the topmos t layers of the MvDN network is robust to view discrepancy, and also discriminat ive. The experiments of face recognition across pose and face recognition across feature type on three datasets with 13 and 2 views respectively demonstrate the superiority of the proposed method, especially compared to the typical linear o nes.
****

Sparsifying Neural Network Connections for Face Recognition
Yi Sun, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Comput er Vision and Pattern Recognition (CVPR), 2016, pp. 4856-4864
This paper proposes to learn high-performance deep ConvNets with sparse neural c onnections, referred to as sparse ConvNets, for face recognition. The sparse Con vNets are learned in an iterative way, each time one additional layer is sparsif ied and the entire model is re-trained given the initial weights learned in prev ious iterations. One important finding is that directly training the sparse Conv Net from scratch failed to find good solutions for face recognition, while using a previously learned denser model to properly initialize a sparser model is cri tical to continue learning effective features for face recognition. This paper a lso proposes a new neural correlation-based weight selection criterion and empir ically verifies its effectiveness in selecting informative connections from prev iously learned models in each iteration. When taking a moderately sparse structu re 26%-76% of weights in the dense model), the proposed sparse ConvNet model sig nificantly improves the face recognition performance of the previous state-of-th e-art DeepID2+ models given the same training data, while it keeps the performan ce of the baseline model with only 12% of the original parameters.
****

Pairwise Linear Regression Classification for Image Set Retrieval
Qingxiang Feng, Yicong Zhou, Rushi Lan; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2016, pp. 4865-4872
This paper proposes the pairwise linear regression classification (PLRC) for ima ge set retrieval. In PLRC, we first define a new concept of the unrelated subspa ce and introduce two strategies to constitute the unrelated subspace. In order t o increase the information of maximizing the query set and the unrelated image s et, we introduce a combination metric for two new classifiers based on two const itution strategies of the unrelated subspace. Extensive experiments on six well- known databases prove that the performance of PLRC is better than that of DLRC a nd several state-of-the-art classifiers for different vision recognition tasks: cluster-based face recognition, video-based face recognition, object recognition and action recognition.
****

The MegaFace Benchmark: 1 Million Faces for Recognition at Scale
Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, Evan Brossard; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4873-4882
Recent face recognition experiments on a major benchmark LFW show stunning perfo rmance--a number of algorithms achieve near to perfect score, surpassing human r ecognition rates. In this paper, we advocate evaluations at the million scale (L FW includes only 13K photos of 5K people). To this end, we have assembled the Me gaFace dataset and created the first MegaFace challenge. Our dataset includes On e Million photos that capture more than 690K different individuals. The challeng e evaluates performance of algorithms with increasing numbers of "distractors" (going from 10 to 1M) in the gallery set. We present both identification and veri fication performance, evaluate performance with respect to pose and a person's a ge, and compare as a function of training data size (#photos and #people). We report results of state of the art and baseline algorithms. The MegaFace datase t, baseline code, and evaluation scripts, are all publicly released for further experimentations at http://megaface.cs.washington.edu.
****

Learnt Quasi-Transitive Similarity for Retrieval From Large Collections of Faces

Ognjen Arandjelovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4883-4892
We are interested in identity-based retrieval of face sets from large unlabelled collections acquired in uncontrolled environments. Given a baseline algorithm for measuring the similarity of two face sets, the meta-algorithm introduced in this paper seeks to leverage the structure of the data corpus to make the best use of the available baseline. In particular, we show how partial transitivity of inter-personal similarity can be exploited to improve the retrieval of particularly challenging sets which poorly match the query under the baseline measure. We: (i) describe the use of proxy sets as a means of computing the similarity between two sets, (ii) introduce transitivity meta-features based on the similarity of salient modes of appearance variation between sets, (iii) show how quasi-transitivity can be learnt from such features without any labelling or manual intervention, and (iv) demonstrate the effectiveness of the proposed methodology through experiments on the notoriously challenging YouTube database.
****

Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition
Yandong Wen, Zhifeng Li, Yu Qiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4893-4901
While considerable progresses have been made on face recognition, age-invariant face recognition (AIFR) still remains a major challenge in real world applications of face recognition systems. The major difficulty of AIFR arises from the fact that the facial appearance is subject to significant intra-personal changes caused by the aging process over time. In order to address this problem, we propose a novel deep face recognition framework to learn the age-invariant deep face features through a carefully designed CNN model. To the best of our knowledge, this is the first attempt to show the effectiveness of deep CNNs in advancing the state-of-the-art of AIFR. Extensive experiments are conducted on several public domain face aging datasets (MORPH Album2, FGNET, and CACD-VS) to demonstrate the effectiveness of the proposed model over the state-of-the-art. We also verify the excellent generalization of our new model on the famous LFW dataset.
****

Copula Ordinal Regression for Joint Estimation of Facial Action Unit Intensity
Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4902-4910
Joint modeling of the intensity of facial action units (AUs) from face images is challenging due to the large number of AUs (30+) and their intensity levels (6). This is in part due to the lack of suitable models that can efficiently handle such a large number of outputs/classes simultaneously, but also due to the lack of target data. For this reason, majority of the methods proposed resort to independent classifiers for the AU intensity. This is suboptimal for at least two reasons: the facial appearance of some AUs changes depending on the intensity of other AUs, and some AUs co-occur more often than others. Encoding this is expected to improve the estimation of target AU intensities, especially in the case of noisy image features, head-pose variations and imbalanced training data. To this end, we introduce a novel modeling framework, Copula Ordinal Regression (COR), that leverages the power of copula functions and CRFs, to detangle the probabilistic modeling of AU dependencies from the marginal modeling of the AU intensity. Consequently, the COR model achieves the joint learning and inference of intensities of multiple AUs, while being computationally tractable. We show on two challenging datasets of naturalistic facial expressions that the proposed approach consistently outperforms (i) independent modeling of AU intensities, and (ii) the state-of-the-art approach for the target task.
****

A Robust Multilinear Model Learning Framework for 3D Faces
Timo Bolkart, Stefanie Wuhrer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4911-4919
Multilinear models are widely used to represent the statistical variations of 3D

human faces as they decouple shape changes due to identity and expression. Existing methods to learn a multilinear face model degrade if not every person is captured in every expression, if face scans are noisy or partially occluded, if expressions are erroneously labeled, or if the vertex correspondence is inaccurate. These limitations impose requirements on the training data that disqualify large amounts of available 3D face data from being usable to learn a multilinear model. To overcome this, we introduce the first framework to robustly learn a multilinear model from 3D face databases with missing data, corrupt data, wrong semantic correspondence, and inaccurate vertex correspondence. To achieve this robustness to erroneous training data, our framework jointly learns a multilinear model and fixes the data. We evaluate our framework on two publicly available 3D face databases, and show that our framework achieves a data completion accuracy that is comparable to state-of-the-art tensor completion methods. Our method reconstructs corrupt data more accurately than state-of-the-art methods, and improves the quality of the learned model significantly for erroneously labeled expressions.

****

Ordinal Regression With Multiple Output CNN for Age Estimation

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4920-4928

To address the non-stationary property of aging patterns, age estimation can be cast as an ordinal regression problem. However, the processes of extracting features and learning a regression model are often separated and optimized independently in previous work. In this paper, we propose an End-to-End learning approach to address ordinal regression problems using deep Convolutional Neural Network, which could simultaneously conduct feature learning and regression modeling. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems. And we propose a multiple output CNN learning algorithm to collectively solve these classification sub-problems, so that the correlation between these tasks could be explored. In addition, we publish an Asian Face Age Dataset (AFAD) containing more than 160K facial images with precise age ground-truths, which is the largest public age dataset to date. To the best of our knowledge, this is the first work to address ordinal regression problems by using CNN, and achieves the state-of-the-art performance on both the MORPH and AFAD datasets.

****

DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4929-4937

This paper considers the task of articulated human pose estimation of multiple people in real world images. We propose an approach that jointly solves the tasks of detection and pose estimation: it infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. This joint formulation is in contrast to previous strategies, that address the problem by first detecting people and subsequently estimating their body pose. We propose a partitioning and labeling formulation of a set of body-part hypotheses generated with CNN-based part detectors. Our formulation, an instance of an integer linear program, implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. Experiments on four different datasets demonstrate state-of-the-art results for both single person and multi person pose estimation.

****

Thin-Slicing for Pose: Learning to Understand Pose Without Explicit Pose Estimation

Suha Kwak, Minsu Cho, Ivan Laptev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4938-4947

We address the problem of learning a pose-aware, compact embedding that projects images with similar human poses to be placed close-by in the embedding space. T

he embedding function is built on a deep convolutional network, and trained with triplet-based rank constraints on real image data. This architecture allows us to learn a robust representation that captures differences in human poses by effectively factoring out variations in clothing, background, and imaging conditions in the wild. For a variety of pose-related tasks, the proposed pose embedding provides a cost-efficient and natural alternative to explicit pose estimation, circumventing challenges of localizing body joints. We demonstrate the efficacy of the embedding on pose-based image retrieval and action recognition problems.
****

A Dual-Source Approach for 3D Pose Estimation From a Single Image
Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, Juergen Gall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4948-4956
One major challenge for 3D pose estimation from a single RGB image is the acquisition of sufficient training data. In particular, collecting large amounts of training data that contain unconstrained images and are annotated with accurate 3D poses is infeasible. We therefore propose to use two independent training sources. The first source consists of images with annotated 2D poses and the second source consists of accurate 3D motion capture data. To integrate both sources, we propose a dual-source approach that combines 2D pose estimation with efficient and robust 3D pose retrieval. In our experiments, we show that our approach achieves state-of-the-art results and is even competitive when the skeleton structure of the two sources differ substantially.
****

Efficiently Creating 3D Training Data for Fine Hand Pose Estimation
Markus Oberweger, Gernot Riegler, Paul Wohlhart, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4957-4965
While many recent hand pose estimation methods critically rely on a training set of labelled frames, the creation of such a dataset is a challenging task that has been overlooked so far. As a result, existing datasets are limited to a few sequences and individuals, with limited accuracy, and this prevents these methods from delivering their full potential. We propose a semi-automated method for efficiently and accurately labeling each frame of a hand depth video with the corresponding 3D locations of the joints: The user is asked to provide only an estimate of the 2D reprojections of the visible joints in some reference frames, which are automatically selected to minimize the labeling work by efficiently optimizing a sub-modular loss function. We then exploit spatial, temporal, and appearance constraints to retrieve the full 3D poses of the hand over the complete sequence. We show that this data can be used to train a recent state-of-the-art hand pose estimation method, leading to increased accuracy.
****

Sparseness Meets Deepness: 3D Human Pose Estimation From Monocular Video
Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4966-4975
This paper addresses the challenge of 3D full-body human pose estimation from a monocular image sequence. Here, two cases are considered: (i) the image locations of the human joints are provided and (ii) the image locations of joints are unknown. In the former case, a novel approach is introduced that integrates a sparsity-driven 3D geometric prior and temporal smoothness. In the latter case, the former case is extended by treating the image locations of the joints as latent variables in order to take into account considerable uncertainties in 2D joint locations. A deep fully convolutional network is trained to predict the uncertainty maps of the 2D joint locations. The 3D pose estimates are realized via an Expectation-Maximization algorithm over the entire sequence, where it is shown that the 2D joint location uncertainties can be conveniently marginalized out during inference. Empirical evaluation on the Human3.6M dataset shows that the proposed approaches achieve greater 3D pose estimation accuracy over state-of-the-art baselines. Further, the proposed approach outperforms a publicly available 2D p

ose estimation baseline on the challenging PennAction dataset.
****

Answer-Type Prediction for Visual Question Answering
Kushal Kafle, Christopher Kanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4976-4984
Recently, algorithms for object recognition and related tasks have become sufficiently proficient that new vision tasks can now be pursued. In this paper, we build a system capable of answering open-ended text-based questions about images, which is known as Visual Question Answering (VQA). Our approach's key insight is that we can predict the form of the answer from the question. We formulate our solution in a Bayesian framework. When our approach is combined with a discriminative model, the combined model achieves state-of-the-art results on four benchmark datasets for open-ended VQA: DAQUAR, COCO-QA, The VQA Dataset, and Visual7W.
****

Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes
Satwik Kottur, Ramakrishna Vedantam, Jose M. F. Moura, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4985-4994
We propose a model to learn visually grounded word embeddings (vis-w2v) to capture visual notions of semantic relatedness. While word embeddings trained using text have been extremely successful, they cannot uncover notions of semantic relatedness implicit in our visual world. For instance, although "eats" and "stares at" seem unrelated in text, they share semantics visually. When people are eating something, they also tend to stare at the food. Grounding diverse relations like "eats" and "stares at" into vision remains challenging, despite recent progress in vision. We note that the visual grounding of words depends on semantics, and not the literal pixels. We thus use abstract scenes created from clipart to provide the visual grounding. We find that the embeddings we learn capture fine-grained, visually grounded notions of semantic relatedness. We show improvements over text-only word embeddings (word2vec) on three tasks: common-sense assertion classification, visual paraphrasing and text-based image retrieval. Our code and datasets are available online.
****

Visual7W: Grounded Question Answering in Images
Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4995-5004
We have seen great progress in basic perceptual tasks such as object recognition and detection. However, AI models still fail to match humans in high-level vision tasks due to the lack of capacities for deeper reasoning. Recently the new task of visual question answering (QA) has been proposed to evaluate a model's capacity for deep image understanding. Previous works have established a loose, global association between QA sentences and images. However, many questions and answers, in practice, relate to local regions in the images. We establish a semantic link between textual descriptions and image regions by object-level grounding. It enables a new type of QA with visual answers, in addition to textual answers used in previous work. We study the visual QA tasks in a grounded setting with a large collection of 7W multiple-choice QA pairs. Furthermore, we evaluate human performance and several baseline models on the QA tasks. Finally, we propose a novel LSTM model with spatial attention to tackle the 7W QA tasks.
****

Learning Deep Structure-Preserving Image-Text Embeddings
Liwei Wang, Yin Li, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5005-5013
This paper proposes a method for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. The network is trained using a large margin objective that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints inspired by metric learning literature. Extensive experiments show that our approach gains significant improvements in accuracy for i

mage-to-text and text-to-image retrieval. Our method achieves new state-of-the-art results on the Flickr30K and MSCOCO image-sentence datasets and shows promise on the new task of phrase localization on the Flickr30K Entities dataset.
****

Yin and Yang: Balancing and Answering Binary Visual Questions
Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5014-5022
The complex compositional structure of language makes problems at the intersection of vision and language challenging. But language also provides a strong prior that can result in good superficial performance, without the underlying models truly understanding the visual content. This can hinder progress in pushing state of art in the computer vision aspects of multi-modal AI. In this paper, we address binary Visual Question Answering (VQA) on abstract scenes. We formulate this problem as visual verification of concepts inquired in the questions. Specifically, we convert the question to a tuple that concisely summarizes the visual concept to be detected in the image. If the concept can be found in the image, the answer to the question is "yes", and otherwise "no". Abstract scenes play two roles (1) They allow us to focus on the high-level semantics of the VQA task as opposed to the low-level recognition problems, and perhaps more importantly, (2) They provide us the modality to balance the dataset such that language priors are controlled, and the role of vision is essential. In particular, we collect fine-grained pairs of scenes for every question, such that the answer to the question is "yes" for one scene, and "no" for the other for the exact same question. Indeed, language priors alone do not perform better than chance on our balanced dataset. Moreover, our proposed approach matches the performance of a state-of-the-art VQA approach on the unbalanced dataset, and outperforms it on the balanced dataset.
****

GIFT: A Real-Time and Scalable 3D Shape Search Engine
Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, Longin Jan Latecki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5023-5032
Projective analysis is an important solution for 3D shape retrieval, since human visual perceptions of 3D shapes rely on various 2D observations from different view points. Although multiple informative and discriminative views are utilized, most projection-based retrieval systems suffer from heavy computational cost, thus cannot satisfy the basic requirement of scalability for search engines. In this paper, we present a real-time 3D shape search engine based on the projective images of 3D shapes. The real-time property of our search engine results from the following aspects: (1) efficient projection and view feature extraction using GPU acceleration; (2) the first inverted file, referred as F-IF, is utilized to speed up the procedure of multi-view matching; (3) the second inverted file (S-IF), which captures a local distribution of 3D shapes in the feature manifold, is adopted for efficient context-based reranking. As a result, for each query the retrieval task can be finished within one second despite the necessary cost of IO overhead. We name the proposed 3D shape search engine, which combines GPU acceleration and Inverted File (Twice), as GIFT. Besides its high efficiency, GIFT also outperforms the state-of-the-art methods significantly in retrieval accuracy on various shape benchmarks and competitions
****

Functional Faces: Groupwise Dense Correspondence Using Functional Maps
Chao Zhang, William A. P. Smith, Arnaud Dessein, Nick Pears, Hang Dai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5033-5041
In this paper we present a method for computing dense correspondence between a set of 3D face meshes using functional maps. The functional maps paradigm brings with it a number of advantages for face correspondence. First, it allows us to combine various notions of correspondence. We do so by proposing a number of face-specific functions, suited to either within- or between-subject correspondence.

Second, we propose a groupwise variant of the method allowing us to compute cycle-consistent functional maps between all faces in a training set. Since functional maps are of much lower dimension than point-to-point correspondences, this is feasible even when the input meshes are very high resolution. Finally, we show how a functional map provides a geometric constraint that can be used to filter feature matches between non-rigidly deforming surfaces.
****

Similarity Metric For Curved Shapes In Euclidean Space
Girum G. Demisse, Djamila Aouada, Bjorn Ottersten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5042-5050
In this paper, we introduce a similarity metric for curved shapes that can be described, distinctively, by ordered points. The proposed method represents a given curve as a point in the deformation space, the direct product of rigid transformation matrices, such that the successive action of the matrices on a fixed starting point reconstructs the full curve. In general, both open and closed curves are represented in the deformation space modulo shape orientation and orientation preserving diffeomorphisms. The use of direct product Lie groups to represent curved shapes led to an explicit formula for geodesic curves and the formulation of a similarity metric between shapes by the L2-norm on the Lie algebra. Additionally, invariance to reparametrization or estimation of point correspondence between shapes is performed as an intermediate step for computing geodesics. Furthermore, since there is no computation of differential quantities on the curves, our representation is more robust to local perturbations and needs no pre-smoothing. We compare our method with the elastic shape metric defined through the square root velocity (SRV) mapping, and other shape matching approaches.
****

Shape Analysis With Hyperbolic Wasserstein Distance
Jie Shi, Wen Zhang, Yalin Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5051-5061
Shape space is an active research field in computer vision study. The shape distance defined in a shape space may provide a simple and refined index to represent a unique shape. Wasserstein distance defines a Riemannian metric for the Wasserstein space. It intrinsically measures the similarities between shapes and is robust to image noise. Thus it has the potential for the 3D shape indexing and classification research. While the algorithms for computing Wasserstein distance have been extensively studied, most of them only work for genus-0 surfaces. This paper proposes a novel framework to compute Wasserstein distance between general topological surfaces with hyperbolic metric. The computational algorithms are based on Ricci flow, hyperbolic harmonic map, and hyperbolic power Voronoi diagram and the method is general and robust. We apply our method to study human facial expression, longitudinal brain cortical morphometry with normal aging, and cortical shape classification in Alzheimer's disease (AD). Experimental results demonstrate that our method may be used as an effective shape index, which outperforms some other standard shape measures in our AD versus healthy control classification study.
****

Tensor Power Iteration for Multi-Graph Matching
Xinchu Shi, Haibin Ling, Weiming Hu, Junliang Xing, Yanning Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5062-5070
Due to its wide range of applications, matching between two graphs has been extensively studied and remains an active topic. By contrast, it is still under-explored on how to jointly match multiple graphs, partly due to its intrinsic computational intractability. In this work, we address this challenging problem in a principled way under the rank-1 tensor approximation framework. In particular, we formulate multi-graph matching as a combinational optimization problem with two main ingredients: unary matching over graph vertices and structure matching over graph edges, both of which across multiple graphs. Then we propose an efficient power iteration solution for the resulted NP-hard optimization problem. The proposed algorithm has several advantages: 1) the intrinsic matching consistency

across multiple graphs based on the high-order tensor optimization; 2) the free employment of powerful high-order node affinity; 3) the flexible integration between various types of node affinities and edge/hyper-edge affinities. Experiments on diverse and challenging datasets validate the effectiveness of the proposed approach in comparison with state-of-the-arts.
****

Multivariate Regression on the Grassmannian for Predicting Novel Domains
Yongxin Yang, Timothy M. Hospedales; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5071-5080
We study the problem of predicting how to recognise visual objects in novel domains with neither labelled nor unlabelled training data. Domain adaptation is now an established research area due to its value in ameliorating the issue of domain shift between train and test data. However, it is conventionally assumed that domains are discrete entities, and that at least unlabelled data is provided in testing domains. In this paper, we consider the case where domains are parametrised by a vector of continuous values (e.g., time, lighting or view angle). We aim to use such domain metadata to predict novel domains for recognition. This allows a recognition model to be pre-calibrated for a new domain in advance (e.g., future time or view angle) without waiting for data collection and re-training. We achieve this by posing the problem as one of multivariate regression on the Grassmannian, where we regress a domain's subspace (point on the Grassmannian) against an independent vector of domain parameters. We derive two novel methodologies to achieve this challenging task: a direct kernel regression, and an indirect method with better extrapolation properties. We evaluate our methods on two cross-domain visual recognition benchmarks, where they perform close to the upper bound of full data domain adaptation. This demonstrates that data is not necessary for domain adaptation if a domain can be parametrically described.
****

Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation
Yao-Hung Hubert Tsai, Yi-Ren Yeh, Yu-Chiang Frank Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5081-5090
While domain adaptation (DA) aims to associate the learning tasks across data domains, heterogeneous domain adaptation (HDA) particularly deals with learning from cross-domain data which are of different types of features. In other words, for HDA, data from source and target domains are observed in separate feature spaces and thus exhibit distinct distributions. In this paper, we propose a novel learning algorithm of Cross-Domain Landmark Selection (CDLS) for solving the above task. With the goal of deriving a domain-invariant feature subspace for HDA, our CDLS is able to identify representative cross-domain data, including the unlabeled ones in the target domain, for performing adaptation. In addition, the adaptation capabilities of such cross-domain landmarks can be determined accordingly. This is the reason why our CDLS is able to achieve promising HDA performance when comparing to state-of-the-art HDA methods. We conduct classification experiments using data across different features, domains, and modalities. The effectiveness of our proposed method can be successfully verified.
****

Geospatial Correspondences for Multimodal Registration
Diego Marcos, Raffay Hamid, Devis Tuia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5091-5100
The growing availability of very high resolution (<1 m/pixel) satellite and aerial images has opened up unprecedented opportunities to monitor and analyze the evolution of land-cover and land-use across the world. To do so, images of the same geographical areas acquired at different times and, potentially, with different sensors must be efficiently parsed to update maps and detect land-cover changes. However, a naive transfer of ground truth labels from one location in the source image to the corresponding location in the target image is not generally feasible, as these images are often only loosely registered (with up to +- 50m of non-uniform errors). Furthermore, land-cover changes in an area over time must be taken into account for an accurate ground truth transfer. To tackle these chal

lenges, we propose a mid-level sensor-invariant representation that encodes imag e regions in terms of the spatial distribution of their spectral neighbors. We i ncorporate this representation in a Markov Random Field to simultaneously accoun t for nonlinear mis-registrations and enforce locality priors to find matches be tween multi-sensor images. We show how our approach can be used to assist in sev eral multimodal land-cover update and change detection problems.
****

Constrained Deep Transfer Feature Learning and Its Applications
Yue Wu, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Patt ern Recognition (CVPR), 2016, pp. 5101-5109
Feature learning with deep models has achieved impressive results for both data representation and classification for various vision tasks. Deep feature learnin g, however,typically requires a large amount of training data, which may not be feasible for some application domains. Transfer learning can be one of the appro aches to alleviate this problem by transferring data from data-rich source domai n to data-scarce target domain. Existing transfer learning methods typically per form one-shot transfer learning and often ignore the specific properties that th e transferred data must satisfy. To address these issues, we introduce a constra ined deep transfer feature learning method to perform simultaneous transfer lear ning and feature learning by performing transfer learning in a progressively imp roving feature space iteratively in order to better narrow the gap between the t arget domain and the source domain for effective transfer of the data from sourc e domain to target domain. Furthermore, we propose to exploit the target domain knowledge and incorporate such prior knowledge as constraint during transfer lea rning to ensure that the transferred data satisfies certain properties of the ta rget domain. To demonstrate the effectiveness of the proposed constrained deep transfer feature learning method, we apply it to thermal feature learning for ey e detection by transferring from the visible domain. We also applied the propose d method for cross-view facial expression recognition as a second application. T he experimental results demonstrate the effectiveness of the proposed method for both applications.
****

Deep Canonical Time Warping
George Trigeorgis, Mihalis A. Nicolaou, Stefanos Zafeiriou, Bjorn W. Schuller; P roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV PR), 2016, pp. 5110-5118
Machine learning algorithms for the analysis of time-series often depend on the assumption that the utilised data are temporally aligned. Any temporal discrepan cies arising in the data is certain to lead to ill-generalisable models, which i n turn fail to correctly capture the properties of the task at hand. The tempora l alignment of time-series is thus a crucial challenge manifesting in a multitud e of applications. Nevertheless, the vast majority of algorithms oriented toward s the temporal alignment of time-series are applied directly on the observation space, or utilise simple linear projections. Thus, they fail to capture complex , hierarchical non-linear representations which may prove to be beneficial towar ds the task of temporal alignment, particularly when dealing with multi-modal da ta (e.g., aligning visual and acoustic information). To this end, we present the Deep Canonical Time Warping (DCTW), a method which automatically learns complex non-linear representations of multiple time-series, generated such that (i) the y are highly correlated, and (ii) temporally in alignment. By means of experimen ts on four real datasets, we show that the representations learnt via the propos ed DCTW significantly outperform state-of-the-art methods in temporal alignment, elegantly handling scenarios with highly heterogeneous features, such as the te mporal alignment of acoustic and visual features.
****

Multilinear Hyperplane Hashing
Xianglong Liu, Xinjie Fan, Cheng Deng, Zhujin Li, Hao Su, Dacheng Tao; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 6, pp. 5119-5127
Hashing has become an increasingly popular technique for fast nearest neighbor s

earch in large databases. Despite its successful progress in classic point-to-po
int search, there are few studies regarding point-to-hyperplane search, which ha
s strong practical capabilities of scaling up in many applications like active l
earning with SVMs. Existing hyperplane hashing methods enable the fast search ba
sed on the randomly generated hash codes, but still suffer from a low collision
probability and thus usually require long codes for a satisfying performance. To
 overcome this problem, this paper proposes a multilinear hyperplane hashing tha
t generates a hash bit using multiple linear projections. Our theoretical analys
is shows that as a product of an even number of random linear projections, the m
ultilinear hash function possesses an increasing power of locality sensitivity t
o the hyperplane queries. To leverage its sensitivity to the angle distance, we
further introduce an angular quantization based learning framework for compact m
ultilinear hashing, which considerably boosts the search performance with less h
ash bits. Experiments with applications to large-scale (up to one million) activ
e learning on two datasets demonstrate the overall superiority of the proposed a
pproach.
****

Large Scale Hard Sample Mining With Monte Carlo Tree Search
Olivier Canevet, Francois Fleuret; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2016, pp. 5128-5137
We investigate an efficient strategy to collect false positives from very large
training sets in the context of object detection. Our approach scales up the sta
ndard bootstrapping procedure by using a hierarchical decomposition of an image
collection which reflects the statistical regularity of the detector's responses
. Based on that decomposition, our procedure uses a Monte Carlo Tree Search to
prioritize the sampling toward sub-families of images which have been observed t
o be rich in false positives, while maintaining a fraction of the sampling towar
d unexplored sub-families of images. The resulting procedure increases substanti
ally the proportion of false positive samples among the visited ones compared to
 a naive uniform sampling. We apply experimentally this new procedure to face d
etection with a collection of 100,000 background images and to pedestrian detect
ion with 32,000 images. We show that for two standard detectors, the proposed st
rategy cuts the number of images to visit by half to obtain the same amount of f
alse positives and the same final performance.
****

Multi-Label Ranking From Positive and Unlabeled Data
Atsushi Kanehira, Tatsuya Harada; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2016, pp. 5138-5146
In this paper, we specifically examine the training of a multi-label classifier
from data with incompletely assigned labels. This problem is fundamentally impor
tant in many multi-label applications because it is almost impossible for human
annotators to assign a complete set of labels, although their judgments are reli
able. In other words, a multi-label dataset usually has properties by which (1)
assigned labels are definitely positive and (2) some labels are absent but are s
till considered positive. Such a setting has been studied as a positive and unla
beled (PU) classification problem in a binary setting. We treat incomplete label
 assignment problems as a multi-label PU ranking, which is an extension of class
ical binary PU problems to the well-studied rank-based multi-label classificatio
n. We derive the conditions that should be satisfied to cancel the negative effe
cts of label incompleteness. Our experimentally obtained results demonstrate the
 effectiveness of these conditions.
****

Joint Unsupervised Learning of Deep Representations and Image Clusters
Jianwei Yang, Devi Parikh, Dhruv Batra; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2016, pp. 5147-5156
In this paper, we propose a recurrent framework for joint unsupervised learning
of deep representations and image clusters. In our framework, successive operati
ons in a clustering algorithm are expressed as steps in a recurrent process, sta
cked on top of representations output by a Convolutional Neural Network (CNN). D
uring training, image clusters and representations are updated jointly: image cl

ustering is conducted in the forward pass, while representation learning in the backward pass. Our key idea behind this framework is that good representations are beneficial to image clustering and clustering results provide supervisory signals to representation learning. By integrating two processes into a single model with a unified weighted triplet loss function and optimizing it end-to-end, we can obtain not only more powerful representations, but also more precise image clusters. Extensive experiments show that our method outperforms the state-of-the-art on image clustering across a variety of image datasets. Moreover, the learned representations generalize well when transferred to other tasks.
****

Kernel Sparse Subspace Clustering on Symmetric Positive Definite Manifolds
Ming Yin, Yi Guo, Junbin Gao, Zhaoshui He, Shengli Xie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5157-5164
Sparse subspace clustering (SSC), as one of the most successful subspace clustering methods, has achieved notable clustering accuracy in computer vision tasks. However, SSC applies only to vector data in Euclidean space. As such, there is still no satisfactory approach to solve subspace clustering by self-expressive principle for symmetric positive definite(SPD) matrices which is very useful in computer vision. In this paper, by embedding the SPD matrices into a Reproducing Kernel Hilbert Space (RKHS), a kernel subspace clustering method is constructed on the SPD manifold through an appropriate Log-Euclidean kernel, termed as kernel sparse subspace clustering on the SPD Riemannian manifold (KSSCR). By exploiting the intrinsic Riemannian Geometry within data, KSSCR can effectively characterize the geodesic distance between SPD matrices to uncover the underlying subspace structure. Experimental results on several famous database demonstrate that the proposed method achieves better clustering results than the state-of-the-art approaches.
****

Symmetry reCAPTCHA
Chris Funk, Yanxi Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5165-5174
This is a reaction to the poor performance of symmetry detection algorithms on real-world images, benchmarked since CVPR 2011. Our systematic study reveals significant difference between human labeled (reflection and rotation) symmetries on photos and the output of computer vision algorithms on the same photo set. We exploit this human-machine symmetry perception gap by proposing a novel symmetry-based Turing test. By leveraging a comprehensive user interface, we collected more than 78,000 symmetry labels from 400 Amazon Mechanical Turk raters on nearly 1,000 photos from the Microsoft COCO dataset. Using a set of ground-truth symmetries automatically generated from noisy human labels, the effectiveness of our work is evidenced by a separate test where over 96% success rate is achieved. We demonstrate statistically significant outcomes for using symmetry perception as a powerful, alternative, image-based reCAPTCHA.
****

Unsupervised Learning of Discriminative Attributes and Visual Representations
Chen Huang, Chen Change Loy, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5175-5184
Attributes offer useful mid-level features to interpret visual data. While most attribute learning methods are supervised by costly human-generated labels, we introduce a simple yet powerful unsupervised approach to learn and predict visual attributes directly from data. Given a large unlabeled image collection as input, we train deep Convolutional Neural Networks (CNNs) to output a set of discriminative, binary attributes often with semantic meanings. Specifically, we first train a CNN coupled with unsupervised discriminative clustering, and then use the cluster membership as a soft supervision to discover shared attributes from the clusters while maximizing their separability. The learned attributes are shown to be capable of encoding rich imagery properties from both natural images and contour patches. The visual representations learned in this way are also transferrable to other tasks such as object detection. We show other convincing results

on the related tasks of image retrieval and classification, and contour detecti
on.
****
When VLAD Met Hilbert
Mehrtash Harandi, Mathieu Salzmann, Fatih Porikli; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5185-5194
In many challenging visual recognition tasks where training data is limited, Ve
ctors of Locally Aggregated Descriptors (VLAD) have emerged as powerful image/vi
deo representations that compete with or outperform state-of-the-art approaches.
In this paper, we address two fundamental limitations of VLAD: its requirement
for the local descriptors to have vector form and its restriction to linear cla
ssifiers due to its high-dimensionality. To this end, we introduce a kernelized
version of VLAD. This not only lets us inherently exploit more sophisticated cla
ssification schemes, but also enables us to efficiently aggregate non-vector des
criptors (e.g., manifold-valued data) in the VLAD framework. Furthermore, we pro
pose an approximate formulation that allows us to accelerate the coding process
while still benefiting from the properties of kernel VLAD. Our experiments demon
strate the effectiveness of our approach at handling manifold-valued data, such
as covariance descriptors, on several classification tasks. Our results also ev
idence the benefits of our nonlinear VLAD descriptors against the linear ones in
Euclidean space using several standard benchmark datasets.
****
Approximate Log-Hilbert-Schmidt Distances Between Covariance Operators for Image
Classification
Ha Quang Minh, Marco San Biagio, Loris Bazzani, Vittorio Murino; Proceedings of
the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.
5195-5203
This paper presents a novel framework for visual object recognition using infini
te-dimensional covariance operators of input features, in the paradigm of kernel
methods on infinite-dimensional Riemannian manifolds. Our formulation provides
a rich representation of image features by exploiting their non-linear correlati
ons, using the power of kernel methods and Riemannian geometry. Theoretically, w
e provide an approximate formulation for the Log-Hilbert-Schmidt distance betwee
n covariance operators that is efficient to compute and scalable to large datase
ts. Empirically, we apply our framework to the task of image classification on e
ight different, challenging datasets. In almost all cases, the results obtained
outperform other state of the art methods, demonstrating the competitiveness and
potential of our framework.
****
Subspace Clustering With Priors via Sparse Quadratically Constrained Quadratic P
rogramming
Yongfang Cheng, Yin Wang, Mario Sznaier, Octavia Camps; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5204-521
2
This paper considers the problem of recovering a subspace arrangement from noisy
samples, potentially corrupted with outliers. Our main result shows that this p
roblem can be formulated as a convex semi-definite optimization problem subject
to an additional rank constrain that involves only a very small number of vari
ables. This is established by first reducing the problem to a (generically non-
convex) quadratically constrained quadratic problem and then using its special s
parse structure to find conditions guaranteeing that a suitably built convex rel
axation is indeed exact. When combined with the commonly used nuclear norm rela
xation for rank, the results above lead to computationally efficient algorithms
with optimality guarantees. A salient feature of the proposed approach is its ab
ility to incorporate existing a-priori information about the noise, co-ocurrenc
es, and percentage of outliers. These results are illustrated with several exam
ples where the proposed algorithm is shown to outperform existing approaches.
****
Robust Tensor Factorization With Unknown Noise
Xi'ai Chen, Zhi Han, Yao Wang, Qian Zhao, Deyu Meng, Yandong Tang; Proceedings o

f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p
p. 5213-5221
Because of the limitations of matrix factorization, such as losing spatial struc
ture information, the concept of tensor factorization has been applied for the r
ecovery of a low dimensional subspace from high dimensional visual data. General
ly, the recovery is achieved by minimizing the loss function between the observe
d data and the factorization representation. Under different assumptions of the
noise distribution, the loss functions are in various forms, like L1 and L2 norm
s. However, real data are often corrupted by noise with an unknown distribution.
Then any specific form of loss function for one specific kind of noise often fa
ils to tackle such real data with unknown noise. In this paper, we propose a ten
sor factorization algorithm to model the noise as a Mixture of Gaussians (MoG).
As MoG has the ability of universally approximating any hybrids of continuous di
stributions, our algorithm can effectively recover the low dimensional subspace
from various forms of noisy observations. The parameters of MoG are estimated un
der the EM framework and through a new developed algorithm of weighted low-rank
tensor factorization (WLRTF). The effectiveness of our algorithm are substantiat
ed by extensive experiments on both of synthetic data and real image data.
****

Kernel Approximation via Empirical Orthogonal Decomposition for Unsupervised Fea
ture Learning
Yusuke Mukuta, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2016, pp. 5222-5230
Kernel approximation methods are important tools for various machine learning pr
oblems. There are two major methods used to approximate the kernel function: the
Nystrom method and the random features method. However, the Nystrom method re
quires relatively high-complexity post-processing to calculate a solution and th
e random features method does not provide sufficient generalization performance.
In this paper, we propose a method that has good generalization performance wit
hout high-complexity postprocessing via empirical orthogonal decomposition using
the probability distribution estimated from training data. We provide a bound f
or the approximation error of the proposed method. Our experiments show that the
proposed method is better than the random features method and comparable with t
he Nystrom method in terms of the approximation error and classification accurac
y. We also show that hierarchical feature extraction using our kernel approximat
ion demonstrates better performance than the existing methods.
****

Active Learning for Delineation of Curvilinear Structures
Agata Mosinska-Domanska, Raphael Sznitman, Przemyslaw Glowacki, Pascal Fua; Proc
eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
, 2016, pp. 5231-5239
Many recent delineation techniques owe much of their increased effectiveness to
path classification algorithms that make it possible to distinguish promisin
g paths from others. The downside of this development is that they require ann
otated training data, which is tedious to produce. In this paper, we propose a
n Active Learning approach that considerably speeds up the annotation process
. Unlike standard ones, it takes advantage of the specificities of the delin
eation problem. It operates on a graph and can reduce the training set size b
y up to 80% without compromising the reconstruction quality. We will show
that our approach outperforms conventional ones on various biomedical an
d natural image datasets, thus showing that it is broadly applicable.
****

Recognizing Emotions From Abstract Paintings Using Non-Linear Matrix Completion
Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, Nicu Sebe; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5240-524
8
Advanced computer vision and machine learning techniques tried to automatically
categorize the emotions elicited by abstract paintings with limited success. Sin
ce the annotation of the emotional content is highly resource-consuming, dataset
s of abstract paintings are either constrained in size or partially annotated. C

onsequently, it is natural to address the targeted task within a transductive fr amework. Intuitively, the use of multi-label classification techniques is desira ble so to synergically exploit the relations between multiple latent variables, such as emotional content, technique, author, etc. A very popular approach for t ransductive multi-label recognition under linear classification settings is matr ix completion. In this study we introduce non-linear matrix completion (NLMC), t hus extending classical linear matrix completion techniques to the non-linear ca se. Together with the theory grounding the model, we propose an efficient optimi zation solver. As shown by our extensive experimental validation on two publicly available datasets, NLMC outperforms state-of-the-art methods when recognizing emotions from abstract paintings.
****

Tensor Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization
Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, Shuicheng Yan; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5249-5257
This paper studies the Tensor Robust Principal Component (TRPCA) problem which e xtends the known Robust PCA to the tensor case. Our model is based on a new tens or Singular Value Decomposition (t-SVD) and its induced tensor tubal rank and t ensor nuclear norm. Consider that we have a 3-way tensor X in R^n*n*n_3 such tha t X=L_0+S_0, where L_0 has low tubal rank and S_0 is sparse. Is that possible t o recover both components? In this work, we prove that under certain suitable as sumptions, we can recover both the low-rank and the sparse components exactly by simply solving a convex program whose objective is a weighted combination of t he tensor nuclear norm and the l1-norm, i.e., min L,E s.t. ||L||_*+lambda||E||_ 1 s.t. X=L+E. where lambda=1/sqrtmax(n_1,n_2)n_3. Interestingly, TRPCA involves RPCA as a special case when n_3=1 and thus it is a simple and elegant tensor ext ension of RPCA. Also numerical experiments verify our theory and the application for the image denoising demonstrates the effectiveness of our method.
****

Sliced Wasserstein Kernels for Probability Distributions
Soheil Kolouri, Yang Zou, Gustavo K. Rohde; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5258-5267
Optimal transport distances, otherwise known as Wasserstein distances, have rece ntly drawn ample attention in computer vision and machine learning as powerful d iscrepancy measures for probability distributions. The recent developments on al ternative formulations of the optimal transport have allowed for faster solution s to the problem and have revamped their practical applications in machine learn ing. In this paper, we exploit the widely used kernel methods and provide a fami ly of provably positive definite kernels based on the Sliced Wasserstein distanc e and demonstrate the benefits of these kernels in a variety of learning tasks. Our work provides a new perspective on the application of optimal transport flav ored distances through kernel methods in machine learning tasks.
****

Trace Quotient Meets Sparsity: A Method for Learning Low Dimensional Image Repre sentations
Xian Wei, Hao Shen, Martin Kleinsteuber; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 5268-5277
This paper presents an algorithm that allows to learn low dimensional representa tions of images in an unsupervised manner. The core idea is to combine two crite ria that play important roles in unsupervised representation learning, namely sp arsity and trace quotient. The former is known to be a convenient tool to identi fy underlying factors, and the latter is known as a disentanglement of underlyin g discriminative factors. In this work, we develop a generic cost function for l earning jointly a sparsifying dictionary and a dimensionality reduction transfor mation. It leads to several counterparts of classic low dimensional representati on methods, such as Principal Component Analysis, Local Linear Embedding, and La placian Eigenmap. Our proposed optimisation algorithm leverages the efficiency o f geometric optimisation on Riemannian manifolds and a closed form solution to t

he elastic net problem.
****
Backtracking ScSPM Image Classifier for Weakly Supervised Top-Down Saliency
Hisham Cholakkal, Jubin Johnson, Deepu Rajan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5278-5287
Top-down saliency models produce a probability map that peaks at target locations specified by a task/goal such as object detection. They are usually trained in a supervised setting involving annotations of objects. We propose a weakly supervised top-down saliency framework using only binary labels that indicate the presence/absence of an object in an image. First, the probabilistic contribution of each image patch to the confidence of an ScSPM-based classifier produces a Reverse-ScSPM (R-ScSPM) saliency map. Neighborhood information is then incorporated through a contextual saliency map which is estimated using logistic regression learnt on patches having high R-ScSPM saliency. Both the saliency maps are combined to obtain the final saliency map. We evaluate the performance of the proposed weakly supervised top-down saliency and achieves comparable performance with fully supervised approaches. Experiments are carried out on 5 challenging datasets across 3 different applications.
****
MSR-VTT: A Large Video Description Dataset for Bridging Video and Language
Jun Xu, Tao Mei, Ting Yao, Yong Rui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5288-5296
While there has been increasing interest in the task of describing video with natural language, current computer vision algorithms are still severely limited in terms of the variability and complexity of the videos and their associated language that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on specific fine-grained domains with limited videos and simple descriptions. While researchers have provided several benchmark datasets for image captioning, we are not aware of any large-scale video description dataset with comprehensive categories yet diverse video content. In this paper we present MSR-VTT (standing for "ABC-Video to Text") which is a new large-scale video benchmark for video understanding, especially the emerging task of translating video to text. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current version, MSR-VTT provides 10K web video clips with 38.7 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary. Each clip is annotated with about 20 natural sentences by 1,327 AMT workers. We present a detailed analysis of MSR-VTT in comparison to a complete set of existing datasets, together with a summarization of different state-of-the-art video-to-text approaches. We also provide an extensive evaluation of these approaches on this dataset, showing that the hybrid Recurrent Neural Network-based approach, which combines single-frame and motion representations with soft-attention pooling strategy, yields the best generalization capability on MSR-VTT.
****
NetVLAD: CNN Architecture for Weakly Supervised Place Recognition
Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297-5307
We tackle the problem of large scale visual place recognition, where the task is to quickly and accurately recognize the location of a given query photograph. We present the following three principal contributions. First, we develop a convolutional neural network (CNN) architecture that is trainable in an end-to-end manner directly for the place recognition task. The main component of this architecture, NetVLAD, is a new generalized VLAD layer, inspired by the "Vector of Locally Aggregated Descriptors" image representation commonly used in image retrieval. The layer is readily pluggable into any CNN architecture and amenable to training via backpropagation. Second, we develop a training procedure, based on a new weakly supervised ranking loss, to learn parameters of the architecture in an end-to-end manner from images depicting the same places over time downloaded fro

m Google Street View Time Machine. Finally, we show that the proposed architectu re significantly outperforms non-learnt image representations and off-the-shelf CNN descriptors on two challenging place recognition benchmarks, and improves ov er current state-of-the-art compact image representations on standard image retr ieval benchmarks.
****

Structural-RNN: Deep Learning on Spatio-Temporal Graphs
Ashesh Jain, Amir R. Zamir, Silvio Savarese, Ashutosh Saxena; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5308-5317
Deep Recurrent Neural Network architectures, though remarkably capable at modeli ng sequences, lack an intuitive high-level spatio-temporal structure. That is wh ile many problems in computer vision inherently have an underlying high-level st ructure and can benefit from it. Spatio-temporal graphs are a popular tool for imposing such high-level intuitions in the formulation of real world problems. I n this paper, we propose an approach for combining the power of high-level spati o-temporal graphs and sequence learning success of Recurrent Neural Networks (RN Ns). We develop a scalable method for casting an arbitrary spatio-temporal graph as a rich RNN mixture that is feedforward, fully differentiable, and jointly tr ainable. The proposed method is generic and principled as it can be used for tra nsforming any spatio-temporal graph through employing a certain set of well defi ned steps. The evaluations of the proposed approach on a diverse set of problems , ranging from modeling human motion to object interactions, shows improvement o ver the state-of-the-art with a large margin. We expect this method to empower new approaches to problem formulation through high-level spatio-temporal graphs and Recurrent Neural Networks.
****

Learning to Select Pre-Trained Deep Representations With Bayesian Evidence Frame work
Yong-Deok Kim, Taewoong Jang, Bohyung Han, Seungjin Choi; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5318-5 326
We propose a Bayesian evidence framework to facilitate transfer learning from pr e-trained deep convolutional neural networks (CNNs). Our framework is formulated on top of a least squares SVM (LS-SVM) classifier, which is simple and fast in both training and testing, and achieves competitive performance in practice. The regularization parameters in LS-SVM is estimated automatically without grid sea rch and cross-validation by maximizing evidence, which is a useful measure to se lect the best performing CNN out of multiple candidates for transfer learning; t he evidence is optimized efficiently by employing Aitken's delta-squared process , which accelerates convergence of fixed point update. The proposed Bayesian evi dence framework also provides a good solution to identify the best ensemble of h eterogeneous CNNs through a greedy algorithm. Our Bayesian evidence framework fo r transfer learning is tested on 12 visual recognition datasets and illustrates the state-of-the-art performance consistently in terms of prediction accuracy an d modeling efficiency.
****

Synthesized Classifiers for Zero-Shot Learning
Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, Fei Sha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5327-533 6
Given semantic descriptions of object classes, zero-shot learning aims to accura tely recognize objects of the unseen classes, from which no examples are availab le at the training stage, by associating them to the seen classes, from which la beled examples are provided. We propose to tackle this problem from the perspect ive of manifold learning. Our main idea is to align the semantic space that is d erived from external information to the model space that concerns itself with re cognizing visual features. To this end, we introduce a set of "phantom" object c lasses whose coordinates live in both the semantic space and the model space. Se rving as bases in a dictionary, they can be optimized from labeled data such tha

t the synthesized real object classifiers achieve optimal discriminative perform ance. We demonstrate superior accuracy of our approach over the state of the art on four benchmark datasets for zero-shot learning, including the full ImageNet Fall 2011 dataset with more than 20,000 unseen classes.
****

Semi-Supervised Vocabulary-Informed Learning
Yanwei Fu, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision a nd Pattern Recognition (CVPR), 2016, pp. 5337-5346
Despite significant progress in object categorization, in recent years, a number of important challenges remain; mainly, ability to learn from limited labeled d ata and ability to recognize object classes within large, potentially open, set of labels. Zero-shot learning is one way of addressing these challenges, but it has only been shown to work with limited sized class vocabularies and typically requires separation between supervised and unsupervised classes, allowing former to inform the latter but not vice versa. We propose the notion of semi-supervis ed vocabulary-informed learning to alleviate the above mentioned challenges and address problems of supervised, zero-shot and open set recognition using a unifi ed framework. Specifically, we propose a maximum margin framework for semantic m anifold-based recognition that incorporates distance constraints from (both supe rvised and unsupervised) vocabulary atoms, ensuring that labeled samples are pro jected closest to their correct prototypes, in the embedding space, than to othe rs. We show that resulting model shows improvements in supervised, zero-shot, an d large open set recognition, with up to 310K class vocabulary on AwA and ImageN et datasets.
****

Simultaneous Clustering and Model Selection for Tensor Affinities
Zhuwen Li, Shuoguang Yang, Loong-Fah Cheong, Kim-Chuan Toh; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5347 -5355
Estimating the number of clusters remains a difficult model selection problem. W e consider this problem in the domain where the affinity relations involve group s of more than two nodes. Building on the previous formulation for the pairwise affinity case, we exploit the mathematical structures in the higher order case. We express the original minimal-rank and positive semi-definite (PSD) constraint s in a form amenable for numerical implementation, as the original constraints a re either intractable or even undefined in general in the higher order case. To scale to large problem sizes, we also propose an alternative formulation, so tha t it can be efficiently solved via stochastic optimization in an online fashion. We evaluate our algorithm with different applications to demonstrate its superi ority, and show it can adapt to varying levels of unbalancedness of clusters.
****

Discriminatively Embedded K-Means for Multi-View Clustering
Jinglin Xu, Junwei Han, Feiping Nie; Proceedings of the IEEE Conference on Compu ter Vision and Pattern Recognition (CVPR), 2016, pp. 5356-5364
In real world applications, more and more data, for example, image/video data, a re high dimensional and represented by multiple views which describe different p erspectives of the data. Efficiently clustering such data is a challenge. To add ress this problem, this paper proposes a novel multi-view clustering method call ed Discriminatively Embedded K-Means (DEKM), which embeds the synchronous learni ng of multiple discriminative subspaces into multi-view K-Means clustering to co nstruct a unified framework, and adaptively control the intercoordinations betwe en these subspaces simultaneously. In this framework, we firstly design a weight ed multi-view Linear Discriminant Analysis (LDA), and then develop an unsupervis ed optimization scheme to alternatively learn the common clustering indicator, m ultiple discriminative subspaces and weights for heterogeneous features with con vergence. Comprehensive evaluations on three benchmark datasets and comparisons with several state-of-the-art multi-view clustering algorithms demonstrate the s uperiority of the proposed work.
****

Min Norm Point Algorithm for Higher Order MRF-MAP Inference

Ishant Shanu, Chetan Arora, Parag Singla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5365-5374
Many tasks in computer vision and machine learning can be modelled as the infere nce problems in an MRF-MAP formulation and can be reduced to minimizing a submod ular function. Using higher order clique potentials to model complex dependencie s between pixels improves the performance but the current state of the art infer ence algorithms fail to scale for larger clique sizes. We adapt a well known Min Norm Point algorithm from mathematical optimization literature to exploit the s um of submodular structure found in the MRF-MAP formulation. Unlike some contemp orary methods, we do not make any assumptions (other than submodularity) on the type of the clique potentials. Current state of the art inference algorithms for general submodular function takes many hours for problems with clique size 16, and fail to scale beyond. On the other hand, our algorithm is highly efficient a nd can perform optimal inference in few seconds even on clique size an order of magnitude larger. The proposed algorithm can even scale to clique sizes of many hundreds, unlocking the usage of really large size cliques for MRF-MAP inference problems in computer vision. We demonstrate the efficacy of our approach by exp erimenting on synthetic as well as real datasets.
****

Learning Deep Representation for Imbalanced Classification
Chen Huang, Yining Li, Chen Change Loy, Xiaoou Tang; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5375-5384
Data in vision domain often exhibit highly-skewed class distribution, i.e., most data belong to a few majority classes, while the minority classes only contain a scarce amount of instances. To mitigate this issue, contemporary classificatio n methods based on deep convolutional neural network (CNN) typically follow clas sic strategies such as class re-sampling or cost-sensitive training. In this pap er, we conduct extensive and systematic experiments to validate the effectivenes s of these classic schemes for representation learning on class-imbalanced data. We further demonstrate that more discriminative deep representation can be lear ned by enforcing a deep network to maintain both inter-cluster and inter-class m argins. This tighter constraint effectively reduces the class imbalance inherent in the local data neighborhood. We show that the margins can be easily deployed in standard deep learning framework through quintuplet instance sampling and th e associated triple-header hinge loss. The representation learned by our approac h, when combined with a simple k-nearest neighbor (kNN) algorithm, shows signifi cant improvements over existing methods on both high- and low-level vision class ification tasks that exhibit imbalanced class distribution.
****

Learning Local Image Descriptors With Deep Siamese and Triplet Convolutional Net works by Minimising Global Loss Functions
Vijay Kumar B G, Gustavo Carneiro, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5385-5394
Recent innovations in training deep convolutional neural network (ConvNet) model s have motivated the design of new methods to automatically learn local image de scriptors. The latest deep ConvNets proposed for this task consist of a siamese network that is trained by penalising misclassification of pairs of local image patches. Current results from machine learning show that replacing this siamese by a triplet network can improve the classification accuracy in several problems , but this has yet to be demonstrated for local image descriptor learning. Moreo ver, current siamese and triplet networks have been trained with stochastic grad ient descent that computes the gradient from individual pairs or triplets of loc al image patches, which can make them prone to overfitting. In this paper, we fi rst propose the use of triplet networks for the problem of local image descripto r learning. Furthermore, we also propose the use of a global loss that minimises the overall classification error of all patches present in the training set, wh ich can improve the generalisation capability of the model. Using the UBC benchm ark dataset for comparing local image descriptors, we show that the triplet netw ork produces a more accurate embedding than the siamese network in terms of the UBC dataset errors. Moreover, we also demonstrate that a combination of the trip

let and global losses produces the best embedding in the field, using this triplet network. Finally, we also show that the use of the central-surround siamese network trained with the global loss produces the best result of the field on the UBC dataset.
****

Sparse Coding for Third-Order Super-Symmetric Tensor Descriptors With Application to Texture Recognition
Piotr Koniusz, Anoop Cherian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5395-5403
Super-symmetric tensors - a higher-order extension of scatter matrices - are becoming increasingly popular in machine learning and computer vision for modeling data statistics, co-occurrences, or even as visual descriptors. They were shown recently to outperform second-order approaches, however, the size of these tensors are exponential in the data dimensionality, which is a significant concern. In this paper, we study third-order super-symmetric tensor descriptors in the context of dictionary learning and sparse coding. For this purpose, we propose a novel non-linear third-order texture descriptor. Our goal is to approximate these tensors as sparse conic combinations of atoms from a learned dictionary. Apart from the significant benefits to tensor compression that this framework offers, our experiments demonstrate that the sparse coefficients produced by this scheme lead to better aggregation of high-dimensional data and showcase superior performance on two common computer vision tasks compared to the state of the art.
****

Random Features for Sparse Signal Classification
Jen-Hao Rick Chang, Aswin C. Sankaranarayanan, B. V. K. Vijaya Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5404-5412
Random features is an approach for kernel-based inference on large datasets. In this paper, we derive performance guarantees for random features on signals, like images, that enjoy sparse representations and show that the number of random features required to achieve a desired approximation of the kernel similarity matrix can be significantly smaller for sparse signals. Based on this, we propose a scheme termed compressive random features that first obtains low-dimensional projections of a dataset and, subsequently, derives random features on the low-dimensional projections. This scheme provides significant improvements in signal dimensionality, computational time, and storage costs over traditional random features while enjoying similar theoretical guarantees for achieving inference performance. We support our claims by providing empirical results across many data sets.
****

High-Quality Depth From Uncalibrated Small Motion Clip
Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5413-5421
We propose a novel approach that generates a high-quality depth map from a set of images captured with a small viewpoint variation, namely small motion clip. As opposed to prior methods that recover scene geometry and camera motions using pre-calibrated cameras, we introduce a self-calibrating bundle adjustment tailored for small motion. This allows our dense stereo algorithm to produce a high-quality depth map for the user without the need for camera calibration. In the dense matching, the distributions of intensity profiles are analyzed to leverage the benefit of having negligible intensity changes within the scene due to the minuscule variation in viewpoint. The depth maps obtained by the proposed framework show accurate and extremely fine structures that are unmatched by previous literature under the same small motion configuration.
****

Efficient 3D Room Shape Recovery From a Single Panorama
Hao Yang, Hui Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5422-5430
We propose a method to recover the shape of a 3D room from a full-view indoor pa

norama. Our algorithm can automatically infer a 3D shape from a collection of pa rtially oriented superpixel facets and line segments. The core part of the algor ithm is a constraint graph, which includes lines and superpixels as vertices, an d encodes their geometric relations as edges. A novel approach is proposed to pe rform 3D reconstruction based on the constraint graph by solving all the geometr ic constraints as constrained linear least-squares. The selected constraints use d for reconstruction are identified using an occlusion detection method with a M arkov random field. Experiments show that our method can recover room shapes tha t can not be addressed by previous approaches. Our method is also efficient, tha t is, the inference time for each panorama is less than 1 minute.
****

Structured Prediction of Unobserved Voxels From a Single Depth Image
Michael Firman, Oisin Mac Aodha, Simon Julier, Gabriel J. Brostow; Proceedings o f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p p. 5431-5440
Building a complete 3D model of a scene, given only a single depth image, is und erconstrained. To gain a full volumetric model, one needs either multiple views, or a single view together with a library of unambiguous 3D models that will fit the shape of each individual object in the scene. We hypothesize that objects of dissimilar semantic classes often share similar 3D shape components, enabling a limited dataset to model the shape of a wide range of objects, and hence esti mate their hidden geometry. Exploring this hypothesis, we propose an algorithm t hat can complete the unobserved geometry of tabletop-sized objects, based on a s upervised model trained on already available volumetric elements. Our model maps from a local observation in a single depth image to an estimate of the surface shape in the surrounding neighborhood. We validate our approach both qualitative ly and quantitatively on a range of indoor object collections and challenging re al scenes.
****

HyperDepth: Learning Depth From Structured Light Without Matching
Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, Shahram Izadi; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2016, pp. 5441-5450
Structured light sensors are popular due to their robustness to untextured scene s and multipath. These systems triangulate depth by solving a correspondence pro blem between each camera and projector pixel. This is often framed as a local st ereo matching task, correlating patches of pixels in the observed and reference image. However, this is computationally intensive, leading to reduced depth accu racy and framerate. We contribute an algorithm for solving this correspondence p roblem efficiently, without compromising depth accuracy. For the first time, thi s problem is cast as a classification-regression task, which we solve extremely efficiently using an ensemble of cascaded random forests. Our algorithm scales i n number of disparities, and each pixel can be processed independently, and in p arallel. No matching or even access to the corresponding reference pattern is re quired at runtime, and regressed labels are directly mapped to depth. Our GPU-ba sed algorithm runs at a 1KHz for 1.3MP input/output images, with disparity error of 0.1 subpixels. We show a prototype high framerate depth camera running at 37 5Hz, useful for solving tracking-related problems. We demonstrate our algorithmi c performance, creating high resolution real-time depth maps that surpass the qu ality of current state of the art depth technologies, highlighting quantization- free results with reduced holes, edge fattening and other stereo-based depth art ifacts.
****

SVBRDF-Invariant Shape and Reflectance Estimation From Light-Field Cameras
Ting-Chun Wang, Manmohan Chandraker, Alexei A. Efros, Ravi Ramamoorthi; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 16, pp. 5451-5459
Light-field cameras have recently emerged as a powerful tool for one-shot passiv e 3D shape capture. However, obtaining the shape of glossy objects like metals, plastics or ceramics remains challenging, since standard Lambertian cues like ph

oto-consistency cannot be easily applied. In this paper, we derive a spatially-v
arying (SV)BRDF-invariant theory for recovering 3D shape and reflectance from li
ght-field cameras. Our key theoretical insight is a novel analysis of diffuse pl
us single-lobe SVBRDFs under a light-field setup. We show that, although direct
shape recovery is not possible, an equation relating depths and normals can stil
l be derived. Using this equation, we then propose using a polynomial (quadratic
) shape prior to resolve the shape ambiguity. Once shape is estimated, we can al
so recover the reflectance. We present extensive synthetic data on the entire ME
RL BRDF dataset, as well as a number of real examples to validate the theory, wh
ere we simultaneously recover shape and BRDFs from a single image taken with a L
ytro Illum camera.
****

Semantic 3D Reconstruction With Continuous Regularization and Ray Potentials Usi
ng a Visibility Consistency Constraint
Nikolay Savinov, Christian Hane, Lubor Ladicky, Marc Pollefeys; Proceedings of t
he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.
5460-5469
We propose an approach for dense semantic 3D reconstruction which uses a data te
rm that is defined as potentials over viewing rays, combined with continuous sur
face area penalization. Our formulation is a convex relaxation which we augment
with a crucial non-convex constraint that ensures exact handling of visibility.
To tackle the non-convex minimization problem, we propose a majorize-minimize ty
pe strategy which converges to a critical point. We demonstrate the benefits of
using the non-convex constraint experimentally. For the geometry-only case, we s
et a new state of the art on two datasets of the commonly used Middlebury multi-
view stereo benchmark. Moreover, our general-purpose formulation directly recons
tructs thin objects, which are usually treated with specialized algorithms. A qu
alitative evaluation on the dense semantic 3D reconstruction task shows that we
improve significantly over previous methods.
****

Theory and Practice of Structure-From-Motion Using Affine Correspondences
Carolina Raposo, Joao P. Barreto; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2016, pp. 5470-5478
Affine Correspondences (ACs) are more informative than Point Correspondences (PC
s) that are used as input in mainstream algorithms for Structure-from-Motion (Sf
M). Since ACs enable to estimate models from fewer correspondences, its use can
dramatically reduce the number of combinations during the iterative step of samp
le-and-test that exists in most SfM pipelines. However, using ACs instead of PCs
as input for SfM passes by fully understanding the relations between ACs and mu
lti-view geometry, as well as by establishing practical, effective AC-based algo
rithms. This article is a step forward into this direction, by providing a clear
account about how ACs constrain the two-view geometry, and by proposing new alg
orithms for plane segmentation and visual odometry that compare favourably with
respect to methods relying in PCs.
****

Just Look at the Image: Viewpoint-Specific Surface Normal Prediction for Improve
d Multi-View Reconstruction
Silvano Galliani, Konrad Schindler; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2016, pp. 5479-5487
We present a multi-view reconstruction method that combines conventional multi-v
iew stereo (MVS) with appearance-based normal prediction, to obtain dense and ac
curate 3D surface models. Reliable surface normals reconstructed from multi-view
correspondence serve as training data for a convolutional neural network (CNN),
which predicts continuous normal vectors from raw image patches. By training fr
om known points in the same image, the prediction is specifically tailored to t
he materials and lighting conditions of the particular scene, as well as to the
precise camera viewpoint. It is therefore a lot easier to learn than generic sin
gle-view normal estimation. The estimated normal maps, together with the known d
epth values from MVS, are integrated to dense depth maps, which in turn are fus
ed into a 3D model. Experiments on the DTU dataset show that our method delivers

3D reconstructions with the same accuracy as MVS, but with significantly higher completeness.
****

From Dusk Till Dawn: Modeling in the Dark
Filip Radenovic, Johannes L. Schonberger, Dinghuang Ji, Jan-Michael Frahm, Ondrej Chum, Jiri Matas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5488-5496
Internet photo collections naturally contain a large variety of illumination conditions, with the largest difference between day and night images. Current modeling techniques do not embrace the broad illumination range often leading to reconstruction failure or severe artifacts. We present an algorithm that leverages the appearance variety to obtain more complete and accurate scene geometry along with consistent multi-illumination appearance information. The proposed method relies on automatic scene appearance grouping, which is used to obtain separate dense 3D models. Subsequent model fusion combines the separate models into a complete and accurate reconstruction of the scene. In addition, we propose a method to derive the appearance information for the model under the different illumination conditions, even for scene parts that are not observed under one illumination condition. To achieve this, we develop a cross-illumination color transfer technique. We evaluate our method on a large variety of landmarks from across Europe reconstructed from a database of 7.4M images.
****

Accelerated Generative Models for 3D Point Cloud Data
Benjamin Eckart, Kihwan Kim, Alejandro Troccoli, Alonzo Kelly, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5497-5505
Finding meaningful, structured representations of 3D point cloud data (PCD) has become a core task for spatial perception applications. In this paper we introduce a method for constructing compact generative representations of PCD at multiple levels of detail. As opposed to deterministic structures such as voxel grids or octrees, we propose probabilistic subdivisions of the data through local mixture modeling, and show how these subdivisions can provide a maximum likelihood segmentation of the data. The final representation is hierarchical, compact, parametric, and statistically derived, facilitating run-time occupancy calculations through stochastic sampling. Unlike traditional deterministic spatial subdivision methods, our technique enables dynamic creation of voxel grids according the application's best needs. In contrast to other generative models for PCD, we explicitly enforce sparsity among points and mixtures, a technique which we call expectation sparsification. This leads to a highly parallel hierarchical Expectation Maximization (EM) algorithm well-suited for the GPU and real-time execution. We explore the trade-offs between model fidelity and model size at various levels of detail, our tests showing favorable performance when compared to octree and NDT-based methods.
****

Monocular Depth Estimation Using Neural Regression Forest
Anirban Roy, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5506-5514
This paper presents a novel deep architecture, called neural regression forest (NRF), for depth estimation from a single image. NRF combines random forests and convolutional neural networks (CNNs). Scanning windows extracted from the image represent samples which are passed down the trees of NRF for predicting their depth. At every tree node, the sample is filtered with a CNN associated with that node. Results of the convolutional filtering are passed to left and right children nodes, i.e., corresponding CNNs, with a Bernoulli probability, until the leaves, where depth estimations are made. CNNs at every node are designed to have fewer parameters than seen in recent work, but their stacked processing along a path in the tree effectively amounts to a deeper CNN. NRF allows for parallelizable training of all "shallow" CNNs, and efficient enforcing of smoothness in depth estimation results. Our evaluation on the benchmark Make3D and NYUv2 datasets demonstrates that NRF outperforms the state of the art, and gracefully handles g

radually decreasing training datasets.
****

DeepStereo: Learning to Predict New Views From the World's Imagery
John Flynn, Ivan Neulander, James Philbin, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5515-5524
Deep networks have recently enjoyed enormous success when applied to recognition and classification problems in computer vision [22, 32], but their use in graphics problems has been limited ([23, 7] are notable recent exceptions). In this work, we present a novel deep architecture that per- forms new view synthesis directly from pixels, trained from a large number of posed image sets. In contrast to tradi- tional approaches which consist of multiple complex stages of processing, each of which require careful tuning and can fail in unexpected ways, our system is trained end-to-end. The pixels from neighboring views of a scene are presented to the network which then directly produces the pixels of the unseen view. The benefits of our approach include gen- erality (we only require posed image sets and can easily apply our method to different domains), and high quality results on traditionally difficult scenes. We believe this is due to the end-to-end nature of our system which is able to plausibly generate pixels according to color, depth, and tex- ture priors learnt automatically from the training data. We show view interpolation results on imagery from the KITTI dataset [12], from data from [1] as well as on StreetView images. To our knowledge, our work is the first to apply deep learning to the problem of new view synthesis from sets of real-world, natural imagery.
****

WIDER FACE: A Face Detection Benchmark
Shuo Yang, Ping Luo, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5525-5533
Face detection is one of the most studied topics in the computer vision community. Much of the progresses have been made by the availability of face detection benchmark datasets. We show that there is a gap between current face detection performance and the real world requirements. To facilitate future face detection research, we introduce the WIDER FACE dataset, which is 10 times larger than existing datasets. The dataset contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. Faces in the proposed dataset are extremely challenging due to large variations in scale, pose and occlusion, as shown in Fig. 1. Furthermore, we show that WIDER FACE dataset is an effective training source for face detection. We benchmark several representative detection systems, providing an overview of state-of-the-art performance and propose a solution to deal with large scale variation. Finally, we discuss common failure cases that worth to be further investigated.
****

Situation Recognition: Visual Semantic Role Labeling for Image Understanding
Mark Yatskar, Luke Zettlemoyer, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5534-5542
This paper introduces situation recognition, the problem of producing a concise summary of the situation an image depicts including: (1) the main activity (e.g., clipping), (2) the participating actors, objects, substances, and locations (e.g., man, shears, sheep, wool, and field) and most importantly (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). We use FrameNet, a verb and role lexicon developed by linguists, to define a large space of possible situations and collect a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. We also introduce structured prediction baselines and show that, in activity-centric images, situation-driven prediction of objects and activities outperforms independent object and activity recognition.
****

A 3D Morphable Model Learnt From 10,000 Faces
James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, David Dunawa

y; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5543-5552
We present Large Scale Facial Model (LSFM) -- a 3D Morphable Model (3DMM) automatically constructed from 9,663 distinct facial identities. To the best of our knowledge LSFM is the largest-scale Morphable Model ever constructed, containing statistical information from a huge variety of the human population. To build such a large model we introduce a novel fully automated and robust Morphable Model construction pipeline. The dataset that LSFM is trained on includes rich demographic information about each subject, allowing for the construction of not only a global 3DMM but also models tailored for specific age, gender or ethnicity groups. As an application example, we utilise the proposed model to perform age classification from 3D shape alone. Furthermore, we perform a systematic analysis of the constructed 3DMMs that showcases their quality and descriptive power. The presented extensive qualitative and quantitative evaluations reveal that the proposed 3DMM achieves state-of-the-art results, outperforming existing models by a large margin. Finally, for the benefit of the research community, we make publicly available the source code of the proposed automatic 3DMM construction pipeline. In addition, the constructed global 3DMM and a variety of bespoke models tailored by age, gender and ethnicity are available on application to researchers involved in medically oriented research.
****

Some Like It Hot - Visual Guidance for Preference Prediction
Rasmus Rothe, Radu Timofte, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5553-5561
For people first impressions of someone are of determining importance. They are hard to alter through further information. This begs the question if a computer can reach the same judgement. Earlier research has already pointed out that age, gender, and average attractiveness can be estimated with reasonable precision. We improve the state-of-the-art, but also predict - based on someone's known preferences - how much that particular person is attracted to a novel face. Our computational pipeline comprises a face detector, convolutional neural networks for the extraction of deep features, standard support vector regression for gender, age and facial beauty, and - as the main novelties - visual regularized collaborative filtering to infer inter-person preferences as well as a novel regression technique for handling visual queries without rating history. We validate the method using a very large dataset from a dating site as well as images from celebrities. Our experiments yield convincing results, i.e. we predict 76% of the ratings correctly solely based on an image, and reveal some sociologically relevant conclusions. We also validate our collaborative filtering solution on the standard MovieLens rating dataset, augmented with movie posters, to predict an individual's movie rating. We demonstrate our algorithms on howhot.io which went viral around the Internet with more than 50 million pictures evaluated in the first month.
****

EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild
C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, Aleix M. Martinez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5562-5570
Research in face perception and emotion theory requires very large annotated databases of images of facial expressions of emotion. Annotations should include Action Units (AUs) and their intensities as well as emotion category. This goal cannot be readily achieved manually. Herein, we present a novel computer vision algorithm to annotate a large database of one million images of facial expressions of emotion in the wild (i.e., face images downloaded from the Internet). First, we show that this newly proposed algorithm can recognize AUs and their intensities reliably across databases. To our knowledge, this is the first published algorithm to achieve highly-accurate results in the recognition of AUs and their intensities across multiple databases. Our algorithm also runs in real-time (>30 images/second), allowing it to work with large numbers of images and video sequen

ces. Second, we use WordNet to download 1,000,000 images of facial expressions with associated emotion keywords from the Internet. These images are then automatically annotated with AUs, AU intensities and emotion categories by our algorithm. The result is a highly useful database that can be readily queried using semantic descriptions for applications in computer vision, affective computing, social and cognitive psychology and neuroscience; e.g., "show me all the images with happy faces" or "all images with AU 1 at intensity c."
****

ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching
Shuxin Ouyang, Timothy M. Hospedales, Yi-Zhe Song, Xueming Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5571-5579
We investigate whether it is possible to improve the performance of automated facial forensic sketch matching by learning from examples of facial forgetting over time. Forensic facial sketch recognition is a key capability for law enforcement, but remains an unsolved problem. It is extremely challenging because there are three distinct contributors to the domain gap between forensic sketches and photos: The well studied sketch-photo modality gap, and the less studied gaps due to (i) the forgetting process of the eye-witness and (ii) their inability to elucidate their memory. In this paper we address the memory problem head on by introducing a database of 400 forensic sketches created at different time-delays. Based on this database we build a model to reverse the forgetting process. Surprisingly, we show that it is possible to systematically "un-forget" facial details. Moreover, it is possible to apply this model to dramatically improve forensic sketch recognition in practice: we achieve state of the art results when matching 195 benchmark forensic sketches against corresponding photos and a 10,030 mug shot database.
****

LOMo: Latent Ordinal Model for Facial Analysis in Videos
Karan Sikka, Gaurav Sharma, Marian Bartlett; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5580-5589
We study the problem of facial analysis in videos. Our first contribution is a novel weakly supervised learning method that models the video event (pain, expression etc.) as a sequence of automatically mined, discriminative sub-events (eg. neutral face, raising brows, contracting lips). The proposed model is inspired by the recent works on Multiple Instance Learning and latent SVM/HCRF- it extends such frameworks to model the ordinal or temporal aspect in the videos, approximately. We show consistent improvements over relevant competitive baselines on four challenging and publicly available video based facial analysis datasets for prediction of expression, clinical pain and intent in dyadic conversations. In combination with complimentary features, we report state-of-the-art results on these datasets.
****

Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation
Dipan K. Pal, Felix Juefei-Xu, Marios Savvides; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5590-5599
We propose an explicitly discriminative and `simple' approach to generate invariance to nuisance transformations modeled as unitary. In practice, the approach works well to handle non-unitary transformations as well. Our theoretical results extend the reach of a recent theory of invariance to discriminative and kernelized features based on unitary kernels. As a special case, a single common framework can be used to generate subject-specific pose-invariant features for face recognition and vice-versa for pose estimation. We show that our main proposed method (DIKF) can perform well under very challenging large-scale semi-synthetic face matching and pose estimation protocols with unaligned faces using no land-marking whatsoever. We additionally benchmark on CMU MPIE and outperform previous work in almost all cases on off-angle face matching while we are on par with the previous state-of-the-art on the LFW unsupervised and image-restricted protocols, without any low-level image descriptors other than raw-pixels.

****
Bottom-Up and Top-Down Reasoning With Hierarchical Rectified Gaussians
Peiyun Hu, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5600-5609
Convolutional neural nets (CNNs) have demonstrated remarkable performance in recent history. Such approaches tend to work in a "unidirectional" bottom-up feed-forward fashion. However, practical experience and biological evidence tells us that feedback plays a crucial role, particularly for detailed spatial understanding tasks. This work explores "bidirectional" architectures that also reason with top-down feedback: neural units are influenced by both lower and higher-level units. We do so by treating units as rectified latent variables in a quadratic energy function, which can be seen as a hierarchical Rectified Gaussian model (RGs). We show that RGs can be optimized with a quadratic program (QP), that can in turn be optimized with a recurrent neural network (with rectified linear units). This allows RGs to be trained with GPU-optimized gradient descent. From a theoretical perspective, RGs help establish a connection between CNNs and hierarchical probabilistic models. From a practical perspective, RGs are well suited for detailed spatial tasks that can benefit from top-down reasoning. We illustrate them on the challenging task of keypoint localization under occlusions, where local bottom-up evidence may be misleading. We demonstrate state-of-the-art results on challenging benchmarks.
****
Fits Like a Glove: Rapid and Reliable Hand Shape Personalization
David Joseph Tan, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, Jamie Shotton; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5610-5619
We present a fast, practical method for personalizing a hand shape basis to an individual user's detailed hand shape using only a small set of depth images. To achieve this, we minimize an energy based on a sum of render-and-compare cost functions called the golden energy. However, this energy is only piecewise continuous, due to pixels crossing occlusion boundaries, and is therefore not obviously amenable to efficient gradient-based optimization. A key insight is that the energy is the combination of a smooth low-frequency function with a high-frequency, low-amplitude, piecewise continuous function. A central finite difference approximation with a suitable step size can therefore jump over the discontinuities to obtain a good approximation to the energy's low-frequency behavior, allowing efficient gradient-based optimization. Experimental results quantitatively demonstrate for the first time that detailed personalized models improve the accuracy of hand tracking and achieve competitive results in both tracking and model registration.
****
Slicing Convolutional Neural Network for Crowd Video Understanding
Jing Shao, Chen-Change Loy, Kai Kang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5620-5628
Learning and capturing both appearance and dynamic representations are pivotal for crowd video understanding. Convolutional Neural Networks (CNNs) have shown its remarkable potential in learning appearance representations from images. However, the learning of dynamic representation, and how it can be effectively combined with appearance features for video analysis, remains an open problem. In this study, we propose a novel spatio-temporal CNN, named Slicing CNN (S-CNN), based on the decomposition of 3D feature maps into 2D spatio- and 2D temporal-slices representations. The decomposition brings unique advantages: (1) the model is capable of capturing dynamics of different semantic units such as groups and objects, (2) it learns separated appearance and dynamic representations while keeping proper interactions between them, and (3) it exploits the selectiveness of spatial filters to discard irrelevant background clutter for crowd understanding. We demonstrate the effectiveness of the proposed S-CNN model on the WWW crowd video dataset for attribute recognition and observe significant performance improvements to the state-of-the-art methods (62.55% from 51.84% [21]).
****

Linear Shape Deformation Models With Local Support Using Graph-Based Structured Matrix Factorisation
Florian Bernard, Peter Gemmar, Frank Hertel, Jorge Goncalves, Johan Thunberg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5629-5638
Representing 3D shape deformations by high-dimensional linear models has many applications in computer vision and medical imaging. Commonly, using Principal Components Analysis a low-dimensional subspace of the high-dimensional shape space is determined. However, the resulting factors (the most dominant eigenvectors of the covariance matrix) have global support, i.e. changing the coefficient of a single factor deforms the entire shape. Based on matrix factorisation with sparsity and graph-based regularisation terms, we present a method to obtain deformation factors with local support. The benefits include better flexibility and interpretability as well as the possibility of interactively deforming shapes locally. We demonstrate that for brain shapes our method outperforms the state of the art in local support models with respect to generalisation and sparse reconstruction, whereas for body shapes our method gives more realistic deformations.
****

Motion From Structure (MfS): Searching for 3D Objects in Cluttered Point Trajectories
Jayakorn Vongkulbhisal, Ricardo Cabral, Fernando De la Torre, Joao P. Costeira; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5639-5647
Object detection has been a long standing problem in computer vision, and state-of-the-art approaches rely on the use of sophisticated features and/or classifiers. However, these learning-based approaches heavily depend on the quality and quantity of labeled data, and do not generalize well to extreme poses or textureless objects. In this work, we explore the use of 3D shape models to detect objects in videos in an unsupervised manner. We call this problem Motion from Structure (MfS): given a set of point trajectories and a 3D model of the object of interest, find a subset of trajectories that correspond to the 3D model and estimate its alignment (i.e., compute the motion matrix). MfS is related to Structure from Motion (SfM) and motion segmentation problems: unlike SfM, the structure of the object is known but the correspondence between the trajectories and the object is unknown; unlike motion segmentation, the MfS problem incorporates 3D structure, providing robustness to tracking mismatches and outliers. Experiments illustrate how our MfS algorithm outperforms alternative approaches in both synthetic data and real videos extracted from YouTube.
****

Volumetric and Multi-View CNNs for Object Classification on 3D Data
Charles R. Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5648-5656
3D shape models are becoming widely available and easier to capture, making available 3D information crucial for progress in object classification. Current state-of-the-art methods rely on CNNs to address this problem. Recently, we witness two types of CNNs being developed: CNNs based upon volumetric representations versus CNNs based upon multi-view representations. Empirical results from these two types of CNNs exhibit a large gap, indicating that existing volumetric CNN architectures and approaches are unable to fully exploit the power of 3D representations. In this paper, we aim to improve both volumetric CNNs and multi-view CNNs according to extensive analysis of existing approaches. To this end, we introduce two distinct network architectures of volumetric CNNs. In addition, we examine multi-view CNNs, where we introduce multi-resolution filtering in 3D. Overall, we are able to outperform current state-of-the-art methods for both volumetric CNNs and multi-view CNNs. We provide extensive experiments designed to evaluate underlying design choices, thus providing a better understanding of the space of methods available for object classification on 3D data.
****

Detecting Vanishing Points Using Global Image Context in a Non-Manhattan World

Menghua Zhai, Scott Workman, Nathan Jacobs; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5657-5665
We propose a novel method for detecting horizontal vanishing points and the zeni
th vanishing point in man-made environments. The dominant trend in existing meth
ods is to first find candidate vanishing points, then remove outliers by enforci
ng mutual orthogonality. Our method reverses this process: we propose a set of h
orizon line candidates and score each based on the vanishing points it contains.
 A key element of our approach is the use of global image context, extracted wit
h a deep convolutional network, to constrain the set of candidates under conside
ration. Our method does not make a Manhattan-world assumption and can operate ef
fectively on scenes with only a single horizontal vanishing point. We evaluate
our approach on three benchmark datasets and achieve state-of-the-art performanc
e on each. In addition, our approach is significantly faster than the previous b
est method.
****
Learning Weight Uncertainty With Stochastic Gradient MCMC for Shape Classificati
on
Chunyuan Li, Andrew Stevens, Changyou Chen, Yunchen Pu, Zhe Gan, Lawrence Carin;
 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (
CVPR), 2016, pp. 5666-5675
Learning the representation of shape cues in 2D & 3D objects for recognition is
a fundamental task in computer vision. Deep neural networks (DNNs) have shown pr
omising performance on this task. Due to the large variability of shapes, accu
rate recognition relies on good estimates of model uncertainty, ignored in tradi
tional training of DNNs, typically learned via stochastic optimization. This pa
per leverages recent advances in stochastic gradient Markov Chain Monte Carlo (S
G-MCMC) to learn weight uncertainty in DNNs. It yields principled Bayesian inter
pretations for the commonly used Dropout/DropConnect techniques and incorporates
 them into the SG-MCMC framework. Extensive experiments on 2D & 3D shape dataset
s and various DNN models demonstrate the superiority of the proposed approach ov
er stochastic optimization. Our approach yields higher recognition accuracy when
 used in conjunction with Dropout and Batch-Normalization.
****
A Field Model for Repairing 3D Shapes
Duc Thanh Nguyen, Binh-Son Hua, Khoi Tran, Quang-Hieu Pham, Sai-Kit Yeung; Proce
edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
 2016, pp. 5676-5684
This paper proposes a field model for repairing 3D shapes constructed from multi
-view RGB data. Specifically, we represent a 3D shape in a Markov random field (
MRF) in which the geometric information is encoded by random binary variables an
d the appearance information is retrieved from a set of RGB images captured at m
ultiple viewpoints. The local priors in the MRF model capture the local structur
es of object shapes and are learnt from 3D shape templates using a convolutional
 deep belief network. Repairing a 3D shape is formulated as the maximum a poster
iori (MAP) estimation in the corresponding MRF. Variational mean field approxima
tion technique is adopted for the MAP estimation. The proposed method was evalua
ted on both artificial data and real data obtained from reconstruction of practi
cal scenes. Experimental results have shown the robustness and efficiency of the
 proposed method in repairing noisy and incomplete 3D shapes.
****
GOGMA: Globally-Optimal Gaussian Mixture Alignment
Dylan Campbell, Lars Petersson; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2016, pp. 5685-5694
Gaussian mixture alignment is a family of approaches that are frequently used fo
r robustly solving the point-set registration problem. However, since they use l
ocal optimisation, they are susceptible to local minima and can only guarantee l
ocal optimality. Consequently, their accuracy is strongly dependent on the quali
ty of the initialisation. This paper presents the first globally-optimal solutio
n to the 3D rigid Gaussian mixture alignment problem under the L2 distance betwe
en mixtures. The algorithm, named GOGMA, employs a branch-and-bound approach to

search the space of 3D rigid motions SE(3), guaranteeing global optimality regar
dless of the initialisation. The geometry of SE(3) was used to find novel upper
and lower bounds for the objective function and local optimisation was integrate
d into the scheme to accelerate convergence without voiding the optimality guara
ntee. The evaluation empirically supported the optimality proof and showed that
the method performed much more robustly on two challenging datasets than an exis
ting globally-optimal registration solution.
****

Efficient Deep Learning for Stereo Matching
Wenjie Luo, Alexander G. Schwing, Raquel Urtasun; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5695-5703
In the past year, convolutional neural networks have been shown to perform extre
mely well for stereo estimation. However, current architectures rely on siamese
networks which exploit concatenation followed by further processing layers, requ
iring a minute of GPU computation per image pair. In contrast, in this paper we
propose a matching network which is able to produce very accurate results in les
s than a second of GPU computation. Towards this goal, we exploit a product laye
r which simply computes the inner product between the two representations of a s
iamese architecture. We train our network by treating the problem as multi-class
 classification, where the classes are all possible disparities. This allows us
to get calibrated scores, which result in much better matching performance when
compared to existing approaches.
****

Efficient Coarse-To-Fine PatchMatch for Large Displacement Optical Flow
Yinlin Hu, Rui Song, Yunsong Li; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2016, pp. 5704-5712
As a key component in many computer vision systems, optical flow estimation, esp
ecially with large displacements, remains an open problem. In this paper we pres
ent a simple but powerful matching method works in a coarse-to-fine scheme for o
ptical flow estimation. Inspired by the nearest neighbor field (NNF) algorithms,
 our approach, called CPM (Coarse-to-fine PatchMatch), blends an efficient rando
m search strategy with the coarse-to-fine scheme for optical flow problem. Unlik
e existing NNF techniques, which is efficient but the results is often too noisy
 for optical flow caused by the lack of global regularization, we propose a prop
agation step with constrained random search radius between adjacent levels on th
e hierarchical architecture. The resulting correspondences enjoys a built-in smo
othing effect, which is more suited for optical flow estimation than NNF techniq
ues. Furthermore, our approach can also capture the tiny structures with large m
otions which is a problem for traditional coarse-to-fine optical flow algorithms
. Interpolated by an edge-preserving interpolation method (EpicFlow), our method
 outperforms the state of the art on MPI-Sintel and KITTI, and runs much faster
than the competing methods.
****

FANNG: Fast Approximate Nearest Neighbour Graphs
Ben Harwood, Tom Drummond; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2016, pp. 5713-5722
We present a new method for approximate nearest neighbour search on large datase
ts of high dimensional feature vectors, such as SIFT or GIST descriptors. Our a
pproach constructs a directed graph that can be efficiently explored for nearest
 neighbour queries. Each vertex in this graph represents a feature vector from t
he dataset being searched. The directed edges are computed by exploiting the fa
ct that, for these datasets, the intrinsic dimensionality of the local manifold-
like structure formed by the elements of the dataset is significantly lower than
 the embedding space. We also provide an efficient search algorithm that uses t
his graph to rapidly find the nearest neighbour to a query with high probability
.   We show how the method can be adapted to give a strong guarantee of 100% rec
all where the query is within a threshold distance of its nearest neighbour. We
demonstrate that our method is significantly more efficient than existing state
of the art methods. In particular, our GPU implementation can deliver 90% recal
l for queries on a data set of 1 million SIFT descriptors at a rate of over 1.2

million queries per second on a Titan X. Finally we also demonstrate how our method scales to datasets of 5M and 20M entries.
****

Exemplar-Driven Top-Down Saliency Detection via Deep Association
Shengfeng He, Rynson W.H. Lau, Qingxiong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5723-5732
Top-down saliency detection is a knowledge-driven search task. While some previous methods aim to learn this "knowledge" from category-specific data, others transfer existing annotations in a large dataset through appearance matching. In contrast, we propose in this paper a locate-by-exemplar strategy. This approach is challenging, as we only use a few exemplars (up to 4) and the appearances among the query object and the exemplars can be very different. To address it, we design a two-stage deep model to learn the intra-class association between the exemplars and query objects. The first stage is for learning object-to-object association, and the second stage is to learn background discrimination. Extensive experimental evaluations show that the proposed method outperforms different baselines and the category-specific models. In addition, we explore the influence of exemplar properties, in terms of exemplar number and quality. Furthermore, we show that the learned model is a universal model and offers great generalization to unseen objects.
****

Unconstrained Salient Object Detection via Proposal Subset Optimization
Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, Radomir Mech; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5733-5742
We aim at detecting salient objects in unconstrained images. In unconstrained images, the number of salient objects (if any) varies from image to image, and is not given. We present a salient object detection system that directly outputs a compact set of detection windows, if any, for an input image. Our system leverages a Convolutional-Neural-Network model to generate location proposals of salient objects. Location proposals tend to be highly overlapping and noisy. Based on the Maximum a Posteriori principle, we propose a novel subset optimization framework to generate a compact set of detection windows out of noisy proposals. In experiments, we show that our subset optimization formulation greatly enhances the performance of our system, and our system attains 16-34% relative improvement in Average Precision compared with the state-of-the-art on three challenging salient object datasets.
****

Recombinator Networks: Learning Coarse-To-Fine Feature Aggregation
Sina Honari, Jason Yosinski, Pascal Vincent, Christopher Pal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5743-5752
Deep neural networks with alternating convolutional, max-pooling and decimation layers are widely used in state of the art architectures for computer vision. Max-pooling purposefully discards precise spatial information in order to create features that are more robust, and typically organized as lower resolution spatial feature maps. On some tasks, such as whole-image classification, max-pooling derived features are well suited; however, for tasks requiring precise localization, such as pixel level prediction and segmentation, max-pooling destroys exactly the information required to perform well. Precise localization may be preserved by shallow convnets without pooling but at the expense of robustness. Can we have our max-pooled multi-layered cake and eat it too? Several papers have proposed summation and concatenation based methods for combining upsampled coarse, abstract features with finer features to produce robust pixel level predictions. Here we introduce another model --- dubbed Recombinator Networks --- where coarse features inform finer features early in their formation such that finer features can make use of several layers of computation in deciding how to use coarse features. The model is trained once, end-to-end and performs better than summation-based architectures, reducing the error from the previous state of the art on two facial keypoint datasets, AFW and AFLW, by 30% and beating the current state

-of-the-art on 300W without using extra data. We improve performance even furthe
r by adding a denoising prediction model based on a novel convnet formulation.
****

End-To-End Saliency Mapping via Probability Distribution Prediction
Saumya Jetley, Naila Murray, Eleonora Vig; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5753-5761
Most saliency estimation methods aim to explicitly model low-level conspicuity c
ues such as edges or blobs and may additionally incorporate top-down cues using
face or text detection. Data-driven methods for training saliency models using e
ye-fixation data are increasingly popular, particularly with the introduction of
 large-scale datasets and deep architectures. However, current methods in this l
atter paradigm use loss functions designed for classification or regression task
s whereas saliency estimation is evaluated on topographical maps. In this work,
we introduce a new saliency map model which formulates a map as a generalized Be
rnoulli distribution. We then train a deep architecture to predict such maps usi
ng novel loss functions which pair the softmax activation function with measures
 designed to compute distances between probability distributions. We show in ext
ensive experiments the effectiveness of such loss functions over standard ones o
n four public benchmark datasets, and demonstrate improved performance over stat
e-of-the-art saliency methods.
****

A Paradigm for Building Generalized Models of Human Image Perception Through Dat
a Fusion
Shaojing Fan, Tian-Tsong Ng, Bryan L. Koenig, Ming Jiang, Qi Zhao; Proceedings o
f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p
p. 5762-5771
In many sub-fields, researchers collect datasets of human ground truth that are
used to create a new algorithm. For example, in research on image perception, da
tasets have been collected for topics such as what makes an image aesthetic or m
emorable. Despite high costs for human data collection, datasets are infrequentl
y reused beyond their own fields of interest. Moreover, the algorithms built fro
m them are domain-specific (predict a small set of attributes) and usually uncon
nected to one another. In this paper, we present a paradigm for building general
ized and expandable models of human image perception. First, we fuse multiple fr
agmented and partially-overlapping datasets through data imputation. We then cre
ate a theoretically-structured statistical model of human image perception that
is fit to the fused datasets. The resulting model has many advantages. (1) It is
 generalized, going beyond the content of the constituent datasets, and can be e
asily expanded by fusing additional datasets. (2) It provides a new ontology usa
ble as a network to expand human data in a cost-effective way. (3) It can guide
the design of a generalized computational algorithm for multi-dimensional visual
 perception. Indeed, experimental results show that a model-based algorithm outp
erforms state-of-the-art methods on predicting visual sentiment, visual realism
and interestingness. Our paradigm can be used in various visual tasks (e.g., vid
eo summarization).
****

Longitudinal Face Modeling via Temporal Deep Restricted Boltzmann Machines
Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Tien D. Bui; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5772-5780
Modeling the face aging process is a challenging task due to large and non-linea
r variations present in different stages of face development. This paper present
s a deep model approach for face age progression that can efficiently capture th
e non-linear aging process and automatically synthesize a series of age-progress
ed faces in various age ranges. In this approach, we first decompose the long-te
rm age progress into a sequence of short-term changes and model it as a face seq
uence. The Temporal Deep Restricted Boltzmann Machines based age progression mod
el together with the prototype faces are then constructed to learn the aging tra
nsformation between faces in the sequence. In addition, to enhance the wrinkles
of faces in the later age ranges, the wrinkle models are further constructed usi
ng Restricted Boltzmann Machines to capture their variations in different facial

regions. The geometry constraints are also taken into account in the last step for more consistent age-progressed results. The proposed approach is evaluated using various face aging databases, i.e. FG-NET, Cross-Age Celebrity Dataset (CACD) and MORPH, and our collected large-scale aging database named AginG Faces in the Wild (AGFW). In addition, when ground-truth age is not available for input image, our proposed system is able to automatically estimate the age of the input face before aging process is employed.

****

Saliency Unified: A Deep Architecture for Simultaneous Eye Fixation Prediction and Salient Object Segmentation

Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, R. Venkatesh Babu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5781-5790

Human eye fixations often correlate with locations of salient objects in the scene. However, only a handful of approaches have attempted to simultaneously address the related aspects of eye fixations and object saliency. In this work, we propose a deep convolutional neural network (CNN) capable of predicting eye fixations and segmenting salient objects in a unified framework. We design the initial network layers, shared between both the tasks, such that they capture the object level semantics and the global contextual aspects of saliency, while the deeper layers of the network address task specific aspects. In addition, our network captures saliency at multiple scales via inception-style convolution blocks. Our network shows a significant improvement over the current state-of-the-art for both eye fixation prediction and salient object segmentation across a number of challenging datasets.

****

Estimating Correspondences of Deformable Objects "In-The-Wild"

Yuxiang Zhou, Epameinondas Antonakos, Joan Alabort-i-Medina, Anastasios Roussos, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5791-5801

During the past few years we have witnessed the development of many methodologies for building and fitting Statistical Deformable Models (SDMs). The construction of accurate SDMs requires careful annotation of images with regards to a consistent set of landmarks. However, the manual annotation of a large amount of images is a tedious, laborious and expensive procedure. Furthermore, for several deformable objects, e.g. human body, it is difficult to define a consistent set of landmarks, and, thus, it becomes impossible to train humans in order to accurately annotate a collection of images. Nevertheless, for the majority of objects, it is possible to extract the shape by object segmentation or even by shape drawing. In this paper, we show for the first time, to the best of our knowledge, that it is possible to construct SDMs by putting object shapes in dense correspondence. Such SDMs can be built with much less effort for a large battery of objects. Additionally, we show that, by sampling the dense model, a part-based SDM can be learned with its parts being in correspondence. We employ our framework to develop SDMs of human arms and legs, which can be used for the segmentation of the outline of the human body, as well as to provide better and more consistent annotations for body joints.

****

Gravitational Approach for Point Set Registration

Vladislav Golyanik, Sk Aziz Ali, Didier Stricker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5802-5810

In this paper a new astrodynamics inspired rigid point set registration algorithm is introduced -- the Gravitational Approach (GA). We formulate point set registration as a modified N-body problem with additional constraints and obtain an algorithm with unique properties which is fully scalable with the number of processing cores. In GA, a template point set moves in a viscous medium under gravitational forces induced by a reference point set. Pose updates are completed by numerically solving the differential equations of Newtonian mechanics. We discuss techniques for efficient implementation of the new algorithm and evaluate it on several synthetic and real-world scenarios. GA is compared with the widely used

Iterative Closest Point and the state of the art rigid Coherent Point Drift algorithms. Experiments evidence that the new approach is robust against noise and can handle challenging scenarios with structured outliers.
****

Context-Aware Gaussian Fields for Non-Rigid Point Set Registration
Gang Wang, Zhicheng Wang, Yufei Chen, Qiangqiang Zhou, Weidong Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5811-5819
Point set registration (PSR) is a fundamental problem in computer vision and pattern recognition, and it has been successfully applied to many applications. Although widely used, existing PSR methods cannot align point sets robustly under degradations, such as deformation, noise, occlusion, outlier, rotation, and multi-view changes. This paper proposes context-aware Gaussian fields (CA-LapGF) for non-rigid PSR subject to global rigid and local non-rigid geometric constraints, where a laplacian regularized term is added to preserve the intrinsic geometry of the transformed set. CA-LapGF uses a robust objective function and the quasi-Newton algorithm to estimate the likely correspondences, and the non-rigid transformation parameters between two point sets iteratively. The CA-LapGF can estimate non-rigid transformations, which are mapped to reproducing kernel Hilbert spaces, accurately and robustly in the presence of degradations. Experimental results on synthetic and real images reveal that how CA-LapGF outperforms state-of-the-art algorithms for non-rigid PSR.
****

Trust No One: Low Rank Matrix Factorization Using Hierarchical RANSAC
Magnus Oskarsson, Kenneth Batstone, Kalle Astrom; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5820-5829
In this paper we present a system for performing low rank matrix factorization. Low-rank matrix factorization is an essential problem in many areas including computer vision, with applications in e.g. affine structure-from-motion, photometric stereo, and non-rigid structure from motion. We specifically target structured data patterns, with outliers and large amounts of missing data. Using recently developed characterizations of minimal solutions to matrix factorization problems with missing data, we show how these can be used as building blocks in a hierarchical system that performs bootstrapping on all levels. This gives an robust and fast system, with state-of-the-art performance.
****

Relaxation-Based Preprocessing Techniques for Markov Random Field Inference
Chen Wang, Ramin Zabih; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5830-5838
Markov Random Fields (MRFs) are a widely used graphical model, but the inference problem is NP-hard. For first-order MRFs with binary labels, Dead End Elimination (DEE) and QPBO can find the optimal labeling for some variables; the much harder case of larger label sets has been addressed by Kovtun and related methods which impose substantial computational overhead. We describe an efficient algorithm to correctly label a subset of the variables for arbitrary MRFs, with particularly good performance on binary MRFs. We propose a sufficient condition to check if a partial labeling is optimal, which is a generalization of DEE's purely local test. We give a hierarchy of relaxations that provide larger optimal partial labelings at the cost of additional computation. Empirical studies were conducted on several benchmarks, using expansion moves for inference. Our algorithm runs in a few seconds, and improves the speed of MRF inference with expansion moves by a factor of 1.5 to 12.
****

Sparse Coding for Classification via Discrimination Ensemble
Yuhui Quan, Yong Xu, Yuping Sun, Yan Huang, Hui Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5839-5847
Discriminative sparse coding has emerged as a promising technique in image analysis and recognition, which couples the process of classifier training and the process of dictionary learning for improving the discriminability of sparse codes. Many existing approaches consider only a simple single linear classifier whose

discriminative power is rather weak. In this paper, we proposed a discriminative sparse coding method which jointly learns a dictionary for sparse coding and an ensemble classifier for discrimination. The ensemble classifier is composed of a set of linear predictors and constructed via both subsampling on data and subspace projection on sparse codes. The advantages of the proposed method over the existing ones are multi-fold: better discriminability of sparse codes, weaker dependence on peculiarities of training data, and more expressibility of classifier for classification. These advantages are also justified in the experiments, as our method outperformed several state-of-the-art methods in several recognition tasks.
****

Principled Parallel Mean-Field Inference for Discrete Random Fields
Pierre Baque, Timur Bagautdinov, Francois Fleuret, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5848-5857
Mean-field variational inference is one of the most popular approaches to inference in discrete random fields. Standard mean-field optimization is based on coordinate descent and in many situations can be impractical. Thus, in practice, various parallel techniques are used, which either rely on ad hoc smoothing with heuristically set parameters, or put strong constraints on the type of models. In this paper, we propose a novel proximal gradient-based approach to optimizing the variational objective. It is naturally parallelizable and easy to implement. We prove its convergence, and then demonstrate that, in practice, it yields faster convergence and often finds better optima than more traditional mean-field optimization techniques. Moreover, our method is less sensitive to the choice of parameters.
****

Guaranteed Outlier Removal With Mixed Integer Linear Programs
Tat-Jun Chin, Yang Heng Kee, Anders Eriksson, Frank Neumann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5858-5866
The maximum consensus problem is fundamentally important to robust geometric fitting in computer vision. Solving the problem exactly is computationally demanding, and the effort required increases rapidly with the problem size. Although randomized algorithms are much more efficient, the optimality of the solution is not guaranteed. Towards the goal of solving maximum consensus exactly, we present guaranteed outlier removal as a technique to reduce the runtime of exact algorithms. Specifically, before conducting global optimization, we attempt to remove data that are provably true outliers, i.e., those that do not exist in the maximum consensus set. We propose an algorithm based on mixed integer linear programming to perform the removal. The result of our algorithm is a smaller data instance that admits much faster solution by a subsequent exact algorithm, while yielding the same globally optimal result as the original problem. We demonstrate that overall speedups of up to 80% can be achieved on common vision problems.
****

Memory Efficient Max Flow for Multi-Label Submodular MRFs
Thalaiyasingam Ajanthan, Richard Hartley, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5867-5876
Multi-label submodular Markov Random Fields (MRFs) have been shown to be solvable using max-flow based on an encoding of the labels proposed by Ishikawa, in which each variable X_i is represented by l nodes (where l is the number of labels) arranged in a column. However, this method in general requires 2l^2 edges for each pair of neighbouring variables. This makes it inapplicable to realistic problems with many variables and labels, due to excessive memory requirement. In this paper, we introduce a variant of the max-flow algorithm that requires much less storage. Consequently, our algorithm makes it possible to optimally solve multi-label submodular problems involving large numbers of variables and labels on a standard computer.
****

Proximal Riemannian Pursuit for Large-Scale Trace-Norm Minimization
Mingkui Tan, Shijie Xiao, Junbin Gao, Dong Xu, Anton van den Hengel, Qinfeng Shi ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5877-5886
Trace-norm regularization plays an important role in many areas such as machine learning and computer vision. Solving trace-norm regularized Trace-norm regularization plays an important role in many areas such as computer vision and machine learning. When solving general large-scale trace-norm regularized problems, existing methods may be computationally expensive due to many high-dimensional truncated singular value decompositions (SVDs) or the unawareness of matrix ranks. In this paper, we propose a proximal Riemannian pursuit (PRP) paradigm which addresses a sequence of trace-norm regularized subproblems defined on nonlinear matrix varieties. To address the subproblem, we extend the proximal gradient method on vector space to nonlinear matrix varieties, in which the SVDs of intermediate solutions are maintained by cheap low-rank QR decompositions, therefore making the proposed method more scalable. Empirical studies on several tasks, such as matrix completion and low-rank representation based subspace clustering, demonstrate the competitive performance of the proposed paradigms over existing methods.
****
Minimizing the Maximal Rank
Erik Bylow, Carl Olsson, Fredrik Kahl, Mikael Nilsson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5887-5895
In computer vision, many problems can be formulated as finding a low rank approximation of a given measurement matrix. Ideally, if all elements of the measurement matrix are available, this is easily solved in the L2-norm using factorization. However, in practice this is rarely the case. Lately, this problem has been addressed using different approaches, one is to replace the rank term by the convex nuclear norm, another is to derive the convex envelope of the rank term plus a data term. In the latter case, matrices are divided into sub-matrices and the envelope is computed for each sub-block individually. In this paper a new convex envelope is derived which takes all sub-matrices into account simultaneously. This leads to a simpler formulation, using only one parameter, for applications where one seeks low rank approximations of multiple matrices with the same rank. We show in this paper how our general framework can be used for manifold denoising of several images at once, as well as just denoising one image. We get comparable results to other well-known methods and our framework can also be used for other applications such as linear shape models.
****
Solving Temporal Puzzles
Caglayan Dicle, Burak Yilmaz, Octavia Camps, Mario Sznaier; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5896-5905
Many physical phenomena, within short time windows, can be explained by low order differential relations. In a discrete world, these relations can be described using low order difference equations or equivalently low order auto regressive (AR) models. In this paper, based on this intuition, we propose an algorithm for solving time-sort temporal puzzles, defined as scrambled time series that need to be sorted out. We frame this highly combinatorial problem using a mixed-integer semi definite programming formulation and show how to turn it into a mixed-integer linear programming problem by using the recently introduced atomic norm framework. Our experiments show the effectiveness and generality of our approach in different scenarios.
****
Estimating Sparse Signals With Smooth Support via Convex Programming and Block Sparsity
Sohil Shah, Tom Goldstein, Christoph Studer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5906-5915
Conventional algorithms for sparse signal recovery and sparse representation rely on l1-norm regularized variational methods. However, when applied to the reconstruction of sparse images, i.e., images where only a few pixels are non-zero, s

imple l1-norm-based methods ignore poten- tial correlations in the support betwe
en adjacent pixels. In a number of applications, one is interested in images tha
t are not only sparse, but also have a support with smooth (or contiguous) bound
aries. Existing algorithms that take into account such a support structure mostl
y rely on non-convex methods and--as a consequence--do not scale well to high-di
mensional problems and/or do not converge to global optima. In this paper, we ex
plore the use of new block l1-norm regularizers, which enforce image sparsity wh
ile simultaneously promoting smooth support structure. By exploiting the convexi
ty of our regularizers, we develop new computationally-efficient recovery algori
thms that guarantee global optimality. We demonstrate the efficacy of our regula
rizers on a variety of imaging tasks including compressive image recovery, image
 restoration, and robust PCA.
****

TenSR: Multi-Dimensional Tensor Sparse Representation
Na Qi, Yunhui Shi, Xiaoyan Sun, Baocai Yin; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5916-5925
The conventional sparse model relies on data representation in the form of vecto
rs. It represents the vector-valued or vectorized one dimensional (1D) version o
f an signal as a highly sparse linear combination of basis atoms from a large di
ctionary. The 1D modeling, though simple, ignores the inherent structure and bre
aks the local correlation inside multidimensional (MD) signals. It also dramatic
ally increases the demand of memory as well as computational resources especiall
y when dealing with high dimensional signals. In this paper, we propose a new sp
arse model TenSR based on tensor for MD data representation along with the corre
sponding MD sparse coding and MD dictionary learning algorithms. The proposed Te
nSR model is able to well approximate the structure in each mode inherent in MD
signals with a series of adaptive separable structure dictionaries via dictionar
y learning. The proposed MD sparse coding algorithm by proximal method further r
educes the computational cost significantly. Experimental results with real worl
d MD signals, i.e. 3D Multi-spectral images, show the proposed TenSR greatly red
uces both the computational and memory costs with competitive performance in com
parison with the state-of-the-art sparse representation methods. We believe our
proposed TenSR model is a promising way to empower the sparse representation esp
ecially for large scale high order signals.
****

Moral Lineage Tracing
Florian Jug, Evgeny Levinkov, Corinna Blasse, Eugene W. Myers, Bjoern Andres; Pr
oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP
R), 2016, pp. 5926-5935
Lineage tracing, the tracking of living cells as they move and divide, is a cent
ral problem in biological image analysis. Solutions, called lineage forests, are
 key to understanding how the structure of multicellular organisms emerges. We p
ropose an integer linear program (ILP) whose feasible solutions define, for ever
y image in a sequence, a decomposition into cells (segmentation) and, across ima
ges, a lineage forest of cells (tracing). In this ILP, path-cut inequalities enf
orce the morality of lineages, i.e., the constraint that cells do not merge. To
find feasible solutions of this NP-hard problem, with certified bounds to the gl
obal optimum, we define efficient separation procedures and apply these as part
of a branch-and-cut algorithm. To show the effectiveness of this approach, we an
alyze feasible solutions for real microscopy data in terms of bounds and run-tim
e, and by their weighted edit distance to lineage forests traced by humans.
****

Globally Optimal Rigid Intensity Based Registration: A Fast Fourier Domain Appro
ach
Behrooz Nasihatkon, Frida Fejne, Fredrik Kahl; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5936-5944
High computational cost is the main obstacle to adapting globally optimal branch
-and-bound algorithms to intensity-based registration. Existing techniques to sp
eed up such algorithms use a multiresolution pyramid of images and bounds on the
 target function among different resolutions for rigidly aligning two images. In

this paper, we propose a dual algorithm in which the optimization is done in the Fourier domain, and multiple resolution levels are replaced by multiple frequency bands. The algorithm starts by computing the target function in lower frequency bands and keeps adding higher frequency bands until the current subregion is either rejected or divided into smaller areas in a branch and bound manner. Unlike spatial multiresolution approaches, to compute the target function for a wider frequency area, one just needs to compute the target in the residual bands. Therefore, if an area is to be discarded, it performs just enough computations required for the rejection. This property also enables us to use a rather large number of frequency bands compared to the limited number of resolution levels used in the space domain algorithm. Experimental results on real images demonstrate considerable speed gains over the space domain method in most cases.

****

On Benefits of Selection Diversity via Bilevel Exclusive Sparsity

Haichuan Yang, Yijun Huang, Lam Tran, Ji Liu, Shuai Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5945-5954

Sparse feature (dictionary) selection is critical for various tasks in computer vision, machine learning, and pattern recognition to avoid overfitting. While extensive research efforts have been conducted on feature selection using sparsity and group sparsity, we note that there has been a lack of development on applications where there is a particular preference on diversity. That is, the selected features are expected to come from different groups or categories. This diversity preference is motivated from many real-world applications such as advertisement recommendation, privacy image classification, and design of survey. In this paper, we proposed a general bilevel exclusive sparsity formulation to pursue the diversity by restricting the overall sparsity and the sparsity in each group. To solve the proposed formulation that is NP hard in general, a heuristic procedure is proposed. The main contributions in this paper include: 1) A linear convergence rate is established for the proposed algorithm; 2) The provided theoretical error bound improves the approaches such as L_1 norm and L_0 types methods which only use the overall sparsity and the quantitative benefits of using the diversity sparsity is provided. To the best of our knowledge, this is the first work to show the theoretical benefits of using the diversity sparsity; 3) Extensive empirical studies are provided to validate the proposed formulation, algorithm, and theory.

****

Fast Training of Triplet-Based Deep Binary Embedding Networks

Bohan Zhuang, Guosheng Lin, Chunhua Shen, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5955-5964

In this paper, we aim to learn a mapping (or embedding) from images to a compact binary space in which Hamming distances correspond to a ranking measure for the image retrieval task. We make use of a triplet loss because this has been shown to be most effective for ranking problems. How- ever, training in previous works can be prohibitively expensive due to the fact that optimization is directly performed on the triplet space, where the number of possible triplets for training is cubic in the number of training examples. To address this issue, we propose to formulate high-order binary codes learning as a multi-label classification problem by explicitly separating learning into two interleaved stages. To solve the first stage, we design a large-scale high-order binary codes inference algorithm to reduce the high-order objective to a standard binary quadratic problem such that graph cuts can be used to efficiently infer the binary codes which serve as the labels of each training datum. In the second stage we propose to map the original image to compact binary codes via carefully designed deep convolutional neural networks (CNNs) and the hash- ing function fitting can be solved by training binary CNN classifiers. An incremental/interleaved optimization strategy is proffered to ensure that these two steps are interactive with each other during training for better accuracy. We conduct experiments on several benchmark datasets, which demonstrate both improved training time (by as much as two orders of magnitude) as well as producing state-of-the- art hashing for various retrieval

tasks.
****
Marr Revisited: 2D-3D Alignment via Surface Normal Prediction
Aayush Bansal, Bryan Russell, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5965-5974
We introduce an approach that leverages surface normal predictions, along with appearance cues, to retrieve 3D models for objects depicted in 2D still images from a large CAD object library. Critical to the success of our approach is the ability to recover accurate surface normals for objects in the depicted scene. We introduce a skip-network model built on the pre-trained Oxford VGG convolutional neural network for surface normal prediction. Our model achieves state-of-the-art accuracy on the NYUv2 RGB-D dataset for surface normal prediction, and recovers fine object detail compared to previous methods. Furthermore, we develop a two-stream network over the input image and predicted surface normals that jointly learns pose and style for CAD model retrieval. When using the predicted surface normals, our two-stream network matches prior work using surface normals computed from RGB-D images on the task of pose prediction, and achieves state of the art when using RGB-D input. Finally, our two-stream network allows us to retrieve CAD models that better match the style and pose of a depicted object compared with baseline approaches.
****
Recovering the Missing Link: Predicting Class-Attribute Associations for Unsupervised Zero-Shot Learning
Ziad Al-Halah, Makarand Tapaswi, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5975-5984
Collecting training images for all visual categories is not only expensive but also impractical. Zero-shot learning (ZSL), especially using attributes, offers a pragmatic solution to this problem. However, at test time most attribute-based methods require a full description of attribute associations for each unseen class. Providing these associations is time consuming and often requires domain specific knowledge. In this work, we aim to carry out attribute-based zero-shot classification in an unsupervised manner. We propose an approach to learn relations that couples class embeddings with their corresponding attributes. Given only the name of an unseen class, the learned relationship model is used to automatically predict the class-attribute associations. Furthermore, our model facilitates transferring attributes across data sets without additional effort. Integrating knowledge from multiple sources results in a significant additional improvement in performance. We evaluate on two public data sets: Animals with Attributes and aPascal/aYahoo. Our approach outperforms state-of-the-art methods in both predicting class-attribute associations and unsupervised ZSL by a large margin.
****
Fast Zero-Shot Image Tagging
Yang Zhang, Boqing Gong, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5985-5994
The well-known word analogy experiments show that the recent word vectors capture fine-grained linguistic regularities in words by linear vector offsets, but it is unclear how well the simple vector offsets can encode visual regularities over words. We study a particular image-word relevance relation in this paper. Our results tell that, given an image, its relevant tags' word vectors rank ahead of the irrelevant tags' along a principal direction in the word vector space. Inspired by this observation, we propose to solve image tagging by estimating the principal direction for an image. Particularly, we exploit linear mappings and nonlinear deep neural networks to approximate the principal direction from an input image. We arrive at a quite versatile tagging model. It runs fast given a test image, in constant time w.r.t. the training set size. It not only gives rise to superior performance for the conventional tagging task on the NUS-WIDE dataset, but also outperforms competitive baselines on annotating images with previously unseen tags. To this end, we name our approach fast zero-shot image tagging (Fast0Tag) to recognize that it possesses the advantages of both FastTag (Chen et al. 2013) and zero-shot learning.

****
Modality and Component Aware Feature Fusion For RGB-D Scene Classification
Anran Wang, Jianfei Cai, Jiwen Lu, Tat-Jen Cham; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5995-6004
While convolutional neural networks (CNN) have been excellent for object recogni tion, the greater spatial variability in scene images typically meant that the s tandard full-image CNN features are suboptimal for scene classification. In this paper, we investigate a framework allowing greater spatial flexibility, in whic h the Fisher vector (FV) encoded distribution of local CNN features, obtained fr om a multitude of region proposals per image, is considered instead. The CNN fea tures are computed from an augmented pixel-wise representation comprising multip le modalities of RGB, HHA and surface normals, as extracted from RGB-D data. Mor e significantly, we make two postulates: (1) component sparsity --- that only a small variety of region proposals and their corresponding FV GMM components cont ribute to scene discriminability, and (2) modal non-sparsity --- within these di scriminative components, all modalities have important contribution. In our fram ework, these are implemented through regularization terms applying group lasso t o GMM components and exclusive group lasso across modalities. By learning and co mbining regressors for both proposal-based FV features and global CNN features, we were able to achieve state-of-the-art scene classification performance on the SUNRGBD Dataset and NYU Depth Dataset V2.
****
PPP: Joint Pointwise and Pairwise Image Label Prediction
Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, Baoxin Li; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6005 -6013
Pointwise label and Pairwise label are both widely used in computer vision tasks . For example, supervised image classification and annotation approaches use poi ntwise label, while attribute-based image relative learning often adopts pairwis e labels. These two types of labels are often considered independently and most existing efforts utilize them separately. However, pointwise labels in image cla ssification and tag annotation are inherently related to the pairwise labels. Fo r example, an image labeled with "coast" and annotated with "beach, sea, sand, s ky" is more likely to have a higher ranking score in terms of the attribute "ope n"; while "men shoes" ranked highly on the attribute "formal" are likely to be a nnotated with "leather, lace up" than "buckle, fabric". The existence of potent ial relations between pointwise labels and pairwise labels motivates us to fuse them together for jointly addressing related vision tasks. In particular, we pro vide a principled way to capture the relations between class labels, tags and at tributes; and propose a novel framework PPP(Pointwise and Pairwise image label P rediction), which is based on overlapped group structure extracted from the poi ntwise-pairwise-label bipartite graph. With experiments on benchmark datasets, w e demonstrate that the proposed framework achieves superior performance on three vision tasks compared to the state-of-the-art methods.
****
Cataloging Public Objects Using Aerial and Street-Level Images - Urban Trees
Jan D. Wegner, Steven Branson, David Hall, Konrad Schindler, Pietro Perona; Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2016, pp. 6014-6023
Each corner of the inhabited world is imaged from multiple viewpoints with incre asing frequency. Online map services like Google Maps or Here Maps provide direc t access to huge amounts of densely sampled, georeferenced images from street vi ew and aerial perspective. There is an opportunity to design computer vision sys tems that will help us search, catalog and monitor public infrastructure, buildi ngs and artifacts. We explore the architecture and feasibility of such a system. The main technical challenge is combining test time information from multiple v iews of each geographic location (e.g., aerial and street views). We implement t wo modules: det2geo, which detects the set of loca- tions of objects belonging t o a given category, and geo2cat, which computes the fine-grained category of the object at a given location. We introduce a solution that adapts state-of-the-ar

t CNN-based object detectors and classifiers. We test our method on "Pasadena Ur
ban Trees", a new dataset of 80,000 trees with geographic and species annotation
s, and show that combining multiple views significantly improves both tree detec
tion and tree species classification, rivaling human performance.
****

Deep Exemplar 2D-3D Detection by Adapting From Real to Rendered Views
Francisco Massa, Bryan C. Russell, Mathieu Aubry; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6024-6033
This paper presents an end-to-end convolutional neural network (CNN) for 2D-3D e
xemplar detection. We demonstrate that the ability to adapt the features of natu
ral images to better align with those of CAD rendered views is critical to the s
uccess of our technique. We show that the adaptation can be learned by compositi
ng rendered views of textured object models on natural images. Our approach can
be naturally incorporated into a CNN detection pipeline and extends the accuracy
 and speed benefits from recent advances in deep learning to 2D-3D exemplar dete
ction. We applied our method to two tasks: instance detection, where we evaluate
d on the IKEA dataset, and object category detection, where we out-perform Aubry
 et al. for "chair" detection on a subset of the Pascal VOC dataset.
****

Zero-Shot Learning via Joint Latent Similarity Embedding
Ziming Zhang, Venkatesh Saligrama; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2016, pp. 6034-6042
Zero-shot recognition (ZSR) deals with the problem of predicting class labels fo
r target domain instances based on source domain side information (e.g. attribut
es) of unseen classes. We formulate ZSR as a binary prediction problem. Our resu
lting classifier is class-independent. It takes an arbitrary pair of source and
target domain instances as input and predicts whether or not they come from the
same class, i.e. whether there is a match. We model the posterior probability of
 a match since it is a sufficient statistic and propose a latent probabilistic m
odel in this context. We develop a joint discriminative learning framework based
 on dictionary learning to jointly learn the parameters of our model for both do
mains, which ultimately leads to our class-independent classifier. Many of the e
xisting embedding methods can be viewed as special cases of our probabilistic mo
del. On ZSR our method shows 4.90% improvement over the state-of-the-art in accu
racy averaged across four benchmark datasets. We also adapt ZSR method for zero-
shot retrieval and show 22.45% improvement accordingly in mean average precision
 (mAP).
****

CRAFT Objects From Images
Bin Yang, Junjie Yan, Zhen Lei, Stan Z. Li; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6043-6051
Object detection is a fundamental problem in image understanding. One popular so
lution is the R-CNN framework and its fast versions. They decompose the object d
etection problem into two cascaded easier tasks: 1) generating object proposals
from images, 2) classifying proposals into various object categories. Despite th
at we are handling with two relatively easier tasks, they are not solved perfect
ly and there's still room for improvement. In this paper, we push the "divide an
d conquer" solution even further by dividing each task into two sub-tasks. We ca
ll the proposed method "CRAFT" (Cascade Region-proposal-network And FasT-rcnn),
which tackles each task with a carefully designed network cascade. We show that
the cascade structure helps in both tasks: in proposal generation, it provides m
ore compact and better localized object proposals; in object classification, it
reduces false positives (mainly between ambiguous categories) by capturing both
inter- and intra-category variances. CRAFT achieves consistent and considerable
improvement over the state-of-the-art on object detection benchmarks like PASCAL
 VOC 07/12 and ILSVRC.
****