Going Deeper With Convolutions
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9
We propose a deep convolutional neural network architecture codenamed Inception that achieves the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC2014). The main hallmark of this architecture is the improved utilization of the computing resources inside the network. By a carefully crafted design, we increased the depth and width of the network while keeping the computational budget constant. To optimize quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation of this architecture, GoogLeNet, a 22 layers deep network, was used to assess its quality in the context of object detection and classification.
****
Propagated Image Filtering
Jen-Hao Rick Chang, Yu-Chiang Frank Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 10-18
In this paper, we propose the propagation filter as a novel image filtering operator, with the goal of smoothing over neighboring image pixels while preserving image context like edges or textural regions. In particular, our filter does not to utilize explicit spatial kernel functions as bilateral and guided filters do. We will show that our propagation filter can be viewed as a robust estimator, which minimizes the expected difference between the filtered and desirable image outputs. We will also relate propagation filtering to belief propagation, and suggest techniques if further speedup of the filtering process is necessary. In our experiments, we apply our propagation filter to a variety of applications such as image denoising, smoothing, fusion, and high-dynamic-range (HDR) compression. We will show that improved performance over existing image filters can be achieved.
****
Web Scale Photo Hash Clustering on A Single Machine
Yunchao Gong, Marcin Pawlowski, Fei Yang, Louis Brandy, Lubomir Bourdev, Rob Fergus; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 19-27
This paper addresses the problem of clustering a very large number of photos (i.e. hundreds of millions a day) in a stream into millions of clusters. This is particularly important as the popularity of photo sharing websites, such as Facebook, Google, and Instagram. Given large number of photos available online, how to efficiently organize them is an open problem. To address this problem, we propose to cluster the binary hash codes of a large number of photos into binary cluster centers. We present a fast binary k-means algorithm that works directly on the similarity-preserving hashes of images and clusters them into binary centers on which we can build hash indexes to speedup computation. The proposed method is capable of clustering millions of photos on a single machine in a few minutes. We show that this approach is usually several magnitude faster than standard k-means and produces comparable clustering accuracy. In addition, we propose an online clustering method based on binary k-means that is capable of clustering large photo stream on a single machine, and show applications to spam detection and trending photo discovery.
****
Expanding Object Detector's Horizon: Incremental Learning Framework for Object Detection in Videos
Alina Kuznetsova, Sung Ju Hwang, Bodo Rosenhahn, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 28-36
Over the last several years it has been shown that image-based object detectors are sensitive to the training data and often fail to generalize to examples that fall outside the original training sample domain (e.g., videos). A number of

domain adaptation (DA) techniques have been proposed to address this problem. DA approaches are designed to adapt a fixed complexity model to the new (e.g., video) domain. We posit that unlabeled data should not only allow adaptation, but also improve (or at least maintain) performance on the original and other domains by dynamically adjusting model complexity and parameters. We call this notion domain expansion. To this end, we develop a new scalable and accurate incremental object detection algorithm, based on several extensions of large-margin embedding (LME). Our detection model consists of an embedding space and multiple class prototypes in that embedding space, that represent object classes; distance to those prototypes allows us to reason about multi-class detection. By incrementally detecting object instances in video and adding confident detections into the model, we are able to dynamically adjust the complexity of the detector over time by instantiating new prototypes to span all domains the model has seen. We test performance of our approach by expanding an object detector trained on ImageNet to detect objects in ego-centric videos of Activity Daily Living (ADL) dataset and challenging YouTube Objects (YTO) dataset.

****

Supervised Discrete Hashing

Fumin Shen, Chunhua Shen, Wei Liu, Heng Tao Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 37-45

Recently, learning based hashing techniques have attracted broad research interests due to the resulting efficient storage and retrieval of images, videos, documents, etc. However, a major difficulty of learning to hash lies in handling the discrete constraints imposed on the needed hash codes, which typically makes hash optimizations very challenging (NP-hard in general). In this work, we propose a new supervised hashing framework, where the learning objective for hashing is to make the optimal binary hash codes for classification. By introducing an auxiliary variable, we reformulate the objective such that it can be solved substantially efficiently by using a regularization algorithm. One of the key steps in the algorithm is to solve the regularization sub-problem associated with the NP-hard binary optimization. We show that with cyclic coordinate descent, the sub-problem admits an analytical solution. As such, a high-quality discrete solution can eventually be obtained in an efficient computing manner, which enables to tackle massive datasets. We evaluate the proposed approach, dubbed Supervised Discrete Hashing (SDH), on four large image datasets, and demonstrate that SDH outperforms the state-of-the-art hashing methods in large-scale image retrieval.

****

What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions?

Mihir Jain, Jan C. van Gemert, Cees G. M. Snoek; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 46-55

This paper contributes to automatic classification and localization of human actions in video. Whereas motion is the key ingredient in modern approaches, we assess the benefits of having objects in the video representation. Rather than considering a handful of carefully selected and localized objects, we conduct an empirical study on the benefit of encoding 15,000 object categories for action using 6 datasets totaling more than 200 hours of video and covering 180 action classes. Our key contributions are i) the first in-depth study of encoding objects for actions, ii) we show that objects matter for actions, and are often semantically relevant as well. iii) We establish that actions have object preferences. Rather than using all objects, selection is advantageous for action recognition. iv) We reveal that object-action relations are generic, which allows to transferring these relationships from the one domain to the other. And, v) objects, when combined with motion, improve the state-of-the-art for both action classification and localization.

****

Landmarks-Based Kernelized Subspace Alignment for Unsupervised Domain Adaptation

Rahaf Aljundi, Remi Emonet, Damien Muselet, Marc Sebban; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 56-63

Domain adaptation (DA) has gained a lot of success in the recent years in comput

er vision to deal with situations where the learning process has to transfer kno wledge from a source to a target domain. In this paper, we introduce a novel uns upervised DA approach based on both subspace alignment and selection of landmark s similarly distributed between the two domains. Those landmarks are selected so as to reduce the discrepancy between the domains and then are used to non linea rly project the data in the same space where an efficient subspace alignment (in closed-form) is performed. We carry out a large experimental comparison in visu al domain adaptation showing that our new method outperforms the most recent uns upervised DA approaches.
****

Blur Kernel Estimation Using Normalized Color-Line Prior
Wei-Sheng Lai, Jian-Jiun Ding, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 64-72
This paper proposes a single-image blur kernel estimation algorithm that utilize s the normalized color-line prior to restore sharp edges without altering edge s tructures or enhancing noise. The proposed prior is derived from the color-line model, which has been successfully applied to non-blind deconvolution and many c omputer vision problems. In this paper, we show that the original color-line pri or is not effective for blur kernel estimation and propose a normalized color-li ne prior which can better enhance edge contrasts. By optimizing the proposed pri or, our method gradually enhances the sharpness of the intermediate patches with out using heuristic filters or external patch priors. The intermediate patches c an then guide the estimation of the blur kernel. A comprehensive evaluation on a large image deblurring dataset shows that our algorithm achieves the state-of-t he-art results.
****

A Light Transport Model for Mitigating Multipath Interference in Time-of-Flight Sensors
Nikhil Naik, Achuta Kadambi, Christoph Rhemann, Shahram Izadi, Ramesh Raskar, Si ng Bing Kang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 73-81
Continuous-wave Time-of-flight (TOF) range imaging has become a commercially via ble technology with many applications in computer vision and graphics. However, the depth images obtained from TOF cameras contain scene dependent errors due to multipath interference (MPI). Specifically, MPI occurs when multiple optical re flections return to a single spatial location on the imaging sensor. Many prior approaches to rectifying MPI rely on sparsity in optical reflections, which is a n extreme simplification. In this paper, we correct MPI by combining the standar d measurements from a TOF camera with information from direct and global light t ransport. We report results on both simulated experiments and physical experimen ts (using the Kinect sensor). Our results, evaluated against ground truth, demon strate a quantitative improvement in depth accuracy.
****

Traditional Saliency Reloaded: A Good Old Model in New Shape
Simone Frintrop, Thomas Werner, German Martin Garcia; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 82-90
In this paper, we show that the seminal, biologically-inspired saliency model by Itti et al. is still competitive with current state-of-the-art methods for sa lient object segmentation if some important adaptions are made. We show which changes are necessary to achieve high performance, with special emphasis on the scale-space: we introduce a twin pyramid for computing Difference-of-Gaus sians, which enables a flexible center-surround ratio. The resulting system, called VOCUS2, is elegant and coherent in structure, fast, and computes salien cy at the pixel level. It is not only suitable for images with few objects, b ut also for complex scenes as captured by mobile devices. Furthermore, we i ntegrate the saliency system into an object proposal generation framework to o btain segment-based saliency maps and boost the results for salient object seg mentation. We show that our system achieves state-of-the-art performance on a large collection of benchmark data.
****

Automatic Construction Of Robust Spherical Harmonic Subspaces
Patrick Snape, Yannis Panagakis, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 91-100
In this paper we propose a method to automatically recover a class specific low dimensional spherical harmonic basis from a set of in-the-wild facial images. We combine existing techniques for uncalibrated photometric stereo and low rank matrix decompositions in order to robustly recover a combined model of shape and identity. We build this basis without aid from a 3D model and show how it can be combined with recent efficient sparse facial feature localisation techniques to recover dense 3D facial shape. Unlike previous works in the area, our method is very efficient and is an order of magnitude faster to train, taking only a few minutes to build a model with over 2000 images. Furthermore, it can be used for real-time recovery of facial shape.
****

Leveraging Stereo Matching With Learning-Based Confidence Measures
Min-Gyu Park, Kuk-Jin Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 101-109
We propose a new approach to associate supervised learning-based confidence prediction with the stereo matching problem. First of all, we analyze the characteristics of various confidence measures in the regression forest framework to select effective confidence measures using training data. We then train regression forests again to predict the correctness (confidence) of a match by using selected confidence measures. In addition, we present a confidence-based matching cost modulation scheme based on the predicted correctness for improving the robustness and accuracy of various stereo matching algorithms. We apply the proposed scheme to the semi-global matching algorithm to make it robust under unexpected difficulties that can occur in outdoor environments. We verify the proposed confidence measure selection and cost modulation methods through extensive experimentation with various aspects using KITTI and challenging outdoor datasets.
****

Saliency Detection via Cellular Automata
Yao Qin, Huchuan Lu, Yiqun Xu, He Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 110-119
In this paper, we introduce Cellular Automata--a dynamic evolution model to intuitively detect the salient object. First, we construct a background-based map using color and space contrast with the clustered boundary seeds. Then, a novel propagation mechanism dependent on Cellular Automata is proposed to exploit the intrinsic relevance of similar regions through interactions with neighbors. Impact factor matrix and coherence matrix are constructed to balance the influential power towards each cell's next state. The saliency values of all cells will be renovated simultaneously according to the proposed updating rule. It's surprising to find out that parallel evolution can improve all the existing methods to a similar level regardless of their original results. Finally, we present an integration algorithm in the Bayesian framework to take advantage of multiple saliency maps. Extensive experiments on six public datasets demonstrate that the proposed algorithm outperforms state-of-the-art methods.
****

Efficient Sparse-to-Dense Optical Flow Estimation Using a Learned Basis and Layers
Jonas Wulff, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 120-130
We address the elusive goal of estimating optical flow both accurately and efficiently by adopting a sparse-to-dense approach. Given a set of sparse matches, we regress to dense optical flow using a learned set of full-frame basis flow fields. We learn the principal components of natural flow fields using flow computed from four Hollywood movies. Optical flow fields are then compactly approximated as a weighted sum of the basis flow fields. Our new PCA-Flow algorithm robustly estimates these weights from sparse feature matches. The method runs in under 200ms/frame on the MPI-Sintel dataset using a single CPU and is more accurate and significantly faster than popular methods such as LDOF and Classic+NL. For some

applications, however, the results are too smooth. Consequently, we develop a n
ovel sparse layered flow method in which each layer is represented by PCA-Flow.
Unlike existing layered methods, estimation is fast because it uses only sparse
matches. We combine information from different layers into a dense flow field us
ing an image-aware MRF. The resulting PCA-Layers method runs in 3.2s/frame, is s
ignificantly more accurate than PCA-Flow, and achieves state-of-the-art performa
nce in occluded regions on MPI-Sintel.
****

Learning Multiple Visual Tasks While Discovering Their Structure
Carlo Ciliberto, Lorenzo Rosasco, Silvia Villa; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 131-139
Multi-task learning is a natural approach for computer vision applications that
require the simultaneous solution of several distinct but related problems, e.g.
object detection, classification, tracking of multiple agents, or denoising, to
name a few. The key idea is that exploring task relatedness (structure) can lea
d to better performances. In this paper, we propose and study a novel sparse, n
on-parametric approach exploiting the theory of Reproducing Kernel Hilbert Spac
es for vector-valued functions. We develop a suitable regularization framework w
hich can be formulated as a convex optimization problem, and is provably solv
able using an alternating minimization approach. Empirical tests show that the
proposed method compares favorably to state of the art techniques and further a
llows to recover interpretable structures, a problem of interest in its own rig
ht.
****

Projection Metric Learning on Grassmann Manifold With Application to Video Based
Face Recognition
Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 140-149
In video based face recognition, great success has been made by representing vid
eos as linear subspaces, which typically lie in a special type of non-Euclidean
space known as Grassmann manifold. To leverage the kernel-based methods develope
d for Euclidean space, several recent methods have been proposed to embed the Gr
assmann manifold into a high dimensional Hilbert space by exploiting the well es
tablished Project Metric, which can approximate the Riemannian geometry of Grass
mann manifold. Nevertheless, they inevitably introduce the drawbacks from tradit
ional kernel-based methods such as implicit map and high computational cost to t
he Grassmann manifold. To overcome such limitations, we propose a novel method t
o learn the Projection Metric directly on Grassmann manifold rather than in Hilb
ert space. From the perspective of manifold learning, our method can be regarded
as performing a geometry-aware dimensionality reduction from the original Grass
mann manifold to a lower-dimensional, more discriminative Grassmann manifold wh
ere more favorable classification can be achieved. Experiments on several real-w
orld video face datasets demonstrate that the proposed method yields competitive
performance compared with the state-of-the-art algorithms.
****

Structural Sparse Tracking
Tianzhu Zhang, Si Liu, Changsheng Xu, Shuicheng Yan, Bernard Ghanem, Narendra Ah
uja, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and
Pattern Recognition (CVPR), 2015, pp. 150-158
Sparse representation has been applied to visual tracking by finding the best ta
rget candidate with minimal reconstruction error by use of target templates. How
ever, most sparse representation based trackers only consider holistic or local
representations and do not make full use of the intrinsic structure among and in
side target candidates, thereby making the representation less effective when si
milar objects appear or under occlusion. In this paper, we propose a novel Struc
tural Sparse Tracking (SST) algorithm, which not only exploits the intrinsic rel
ationship among target candidates and their local patches to learn their sparse
representations jointly, but also preserves the spatial layout structure among t
he local patches inside each target candidate. We show that our SST algorithm ac
commodates most existing sparse trackers with the respective merits. Both qualit

ative and quantitative evaluations on challenging benchmark image sequences demo nstrate that the proposed SST algorithm performs favorably against several state -of-the-art methods.
****

Data-Driven Depth Map Refinement via Multi-Scale Sparse Representation
HyeokHyen Kwon, Yu-Wing Tai, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 159-167
Depth maps captured by consumer-level depth cameras such as Kinect are usually d egraded by noise, missing values, and quantization. In this paper, we present a data-driven approach for refining degraded RAW depth maps that are coupled with an RGB image. The key idea of our approach is to take advantage of a training se t of high-quality depth data and transfer its information to the RAW depth map t hrough multi-scale dictionary learning. Utilizing a sparse representation, our m ethod learns a dictionary of geometric primitives which captures the correlation between high-quality mesh data, RAW depth maps and RGB images. The dictionary i s learned and applied in a manner that accounts for various practical issues tha t arise in dictionary-based depth refinement. Compared to previous approaches th at only utilize the correlation between RAW depth maps and RGB images, our metho d produces improved depth maps without over-smoothing. Since our approach is dat a driven, the refinement can be targeted to a specific class of objects by emplo ying a corresponding training set. In our experiments, we show that this leads t o additional improvements in recovering depth maps of human faces.
****

Uncalibrated Photometric Stereo Based on Elevation Angle Recovery From BRDF Symm etry of Isotropic Materials
Feng Lu, Imari Sato, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 168-176
This paper addresses the problem of uncalibrated photometric stereo with isotrop ic reflectances. Existing methods face difficulty in solving for the elevation a ngles of surface normals when the light sources only cover the visible hemispher e. Here, we introduce the notion of "constrained half-vector symmetry" for gener al isotropic BRDFs and show its capability of elevation angle recovery. This sor t of symmetry can be observed in a 1D BRDF slice from a subset of surface normal s with the same azimuth angle, and we use it to devise an efficient modeling and solution method to constrain and recover the elevation angles of surface normal s accurately. To enable our method to work in an uncalibrated manner, we further solve for light sources in the case of general isotropic BRDFs. By combining th is method with the existing ones for azimuth angle estimation, we can get state- of-the-art results for uncalibrated photometric stereo with general isotropic re flectances.
****

Attributes and Categories for Generic Instance Search From One Example
Ran Tao, Arnold W.M. Smeulders, Shih-Fu Chang; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 177-186
This paper aims for generic instance search from one example where the instance can be an arbitrary 3D object like shoes, not just near-planar and one-sided ins tances like buildings and logos. Firstly, we evaluate state-of-the-art instance search methods on this problem. We observe that what works for buildings loses i ts generality on shoes. Secondly, we propose to use automatically learned catego ry-specific attributes to address the large appearance variations present in gen eric instance search. On the problem of searching among instances from the same category as the query, the category-specific attributes outperform existing appr oaches by a large margin. On a shoe dataset containing 6624 shoe images recorded from all viewing angles, we improve the performance from 36.73 to 56.56 using c ategory-specific attributes. Thirdly, we extend our methods to search objects wi thout restricting to the specifically known category. We show the combination of category-level information and the category-specific attributes is superior to combining category-level information with low-level features such as Fisher vect or.
****

Heat Diffusion Over Weighted Manifolds: A New Descriptor for Textured 3D Non-Rigid Shapes
Mostafa Abdelrahman, Aly Farag, David Swanson, Moumen T. El-Melegy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 187-195
This paper propose an approach for modeling textured 3D non-rigid models based on Weighted Heat Kernel Signature(W-HKS). As a first contribution, we show how to include photometric information as a weight over the shape manifold, we also propose a novel formulation for heat diffusion over weighted manifolds. As a second contribution we present a new discretization method for the proposed equation using finite element approximation. Finally, the weighted heat kernel signature is used as a shape descriptor. The proposed descriptor encodes both the photometric, and geometric information based on the solution of one equation. We also propose a new method to introduce the scale invariance for the weighted heat kernel signature. The performance is tested on two benchmark datasets. The results have indeed confirmed the high performance of the proposed approach on the textured shape retrieval problem, and show that the proposed method is useful in coping with different challenges of shape analysis where pure geometric and pure photometric methods fail.
****

A Dynamic Programming Approach for Fast and Robust Object Pose Recognition From Range Images
Christopher Zach, Adrian Penate-Sanchez, Minh-Tri Pham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 196-203
Joint object recognition and pose estimation solely from range images is an important task e.g. in robotics applications and in automated manufacturing environments. The lack of color information and limitations of current commodity depth sensors make this task a challenging computer vision problem, and a standard random sampling based approach is prohibitively time-consuming. We propose to address this difficult problem by generating promising inlier sets for pose estimation by early rejection of clear outliers with the help of local belief propagation (or dynamic programming). By exploiting data-parallelism our method is fast, and we also do not rely on a computationally expensive training phase. We demonstrate state-of-the art performance on a standard dataset and illustrate our approach on challenging real sequences.
****

Beyond Gaussian Pyramid: Multi-Skip Feature Stacking for Action Recognition
Zhengzhong Lan, Ming Lin, Xuanchong Li, Alex G. Hauptmann, Bhiksha Raj; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 204-212
Most state-of-the-art action feature extractors involve differential operators, which act as highpass filters and tend to attenuate low frequency action information. This attenuation introduces bias to the resulting features and generates ill-conditioned feature matrices. The Gaussian Pyramid has been used as a feature enhancing technique that encodes scale-invariant characteristics into the feature space in an attempt to deal with this attenuation. However, at the core of the Gaussian Pyramid is a convolutional smoothing operation, which makes it incapable of generating new features at coarse scales. In order to address this problem, we propose a novel feature enhancing technique called Multi-skIp Feature Stacking (MIFS), which stacks features extracted using a family of differential filters parameterized with multiple time skips and encodes shift-invariance into the frequency space. MIFS compensates for information lost from using differential operators by recapturing information at coarse scales. This recaptured information allows us to match actions at different speeds and ranges of motion. We prove that MIFS enhances the learnability of differential-based features exponentially. The resulting feature matrices from MIFS have much smaller conditional numbers and variances than those from conventional methods. Experimental results show significantly improved performance on challenging action recognition and event detection tasks. Specifically, our method exceeds the state-of-the-arts on Hollywood2, UCF101 and UCF50 datasets and is comparable to state-of-the-arts on HMDB

51 and Olympics Sports datasets. MIFS can also be used as a speedup strategy for feature extraction with minimal or no accuracy cost.
****

A Geodesic-Preserving Method for Image Warping
Dongping Li, Kaiming He, Jian Sun, Kun Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 213-221
The manipulation of panoramic/wide-angle images is usually achieved via image warping. Though various techniques have been developed for preserving shapes and straight lines for warping, these are not sufficient for panoramic/wide-angle images. The image projections will turn the straight lines into curved "geodesic lines", and it is fundamentally impossible to keep all these lines straight. In this work, we propose a geodesic-preserving method for content-aware image warping. An energy term is introduced to preserve the geodesic appearance of the geodesic lines, and can be used with shape-preserving terms. Our method is demonstrated in various applications, including rectangling panoramas, resizing panoramic/wide-angle images, and wide-angle image manipulation. An extension to ellipse preservation for general images is also presented.
****

Shape Driven Kernel Adaptation in Convolutional Neural Network for Robust Facial Traits Recognition
Shaoxin Li, Junliang Xing, Zhiheng Niu, Shiguang Shan, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 222-230
One key challenge of facial traits recognition is the large non-rigid appearance variations due to irrelevant real world factors, such as viewpoint and expression changes. In this paper, we explore how the shape information, i.e. facial landmark positions, can be explicitly deployed into the popular Convolutional Neural Network (CNN) architecture to disentangling such irrelevant non-rigid appearance variations. First, instead of using fixed kernels, we propose a kernel adaptation method to dynamically determine the convolutional kernels according to the distribution of facial landmarks, which helps learning more robust features. Second, motivated by the intuition that different local facial regions may demand different adaptation functions, we further propose a tree-structured convolutional architecture to hierarchically fuse multiple local adaptive CNN subnetworks. Comprehensive experiments on WebFace, Morph II and MultiPIE databases well validate the effectiveness of the proposed kernel adaptation method and tree-structured convolutional architecture for facial traits recognition tasks, including identity, age and gender classification. For all the tasks, the proposed architecture consistently achieves the state-of-the-art performances.
****

From Categories to Subcategories: Large-Scale Image Classification With Partial Class Label Refinement
Marko Ristin, Juergen Gall, Matthieu Guillaumin, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 231-239
The number of digital images is growing extremely rapidly, and so is the need for their classification. But, as more images of pre-defined categories become available, they also become more diverse and cover finer semantic differences. Ultimately, the categories themselves need to be divided into subcategories to account for that semantic refinement. Image classification in general has improved significantly over the last few years, but it still requires a massive amount of manually annotated data. Subdividing categories into subcategories multiples the number of labels, aggravating the annotation problem. Hence, we can expect the annotations to be refined only for a subset of the already labeled data, and exploit coarser labeled data to improve classification. In this work, we investigate how coarse category labels can be used to improve the classification of subcategories. To this end, we adopt the framework of Random Forests and propose a regularized objective function that takes into account relations between categories and subcategories. Compared to approaches that disregard the extra coarse labeled data, we achieve a relative improvement in subcategory classificatio

n accuracy of up to 22% in our large-scale image classification experiments.
****

Combination Features and Models for Human Detection
Yunsheng Jiang, Jinwen Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 240-248
This paper presents effective combination models with certain combination features for human detection. In the past several years, many existing features/models have achieved impressive progress, but their performances are still limited by the biases rooted in their self-structures, that is, a particular kind of feature/model may work well for some types of human bodies, but not for all the types. To tackle this difficult problem, we combine certain complementary features/models together with effective organization/fusion methods. Specifically, the HOG features, color features and bar-shape features are combined together with a cell-based histogram structure to form the so-called HOG-III features. Moreover, the detections from different models are fused together with the new proposed weighted-NMS algorithm, which enhances the probable "true" activations as well as suppresses the overlapped detections. The experiments on PASCAL VOC datasets demonstrate that, both the HOG-III features and the weighted-NMS fusion algorithm are effective (obvious improvement for detection performance) and efficient (relatively less computation cost): When applied to human detection task with the Grammar model and Poselet model, they can boost the detection performance significantly; Also, when extended to detection of the whole VOC 20 object categories with the deformable part-based model and deepCNN-based model, they still show competitive improvements.
****

Improving Object Detection With Deep Convolutional Networks via Bayesian Optimization and Structured Prediction
Yuting Zhang, Kihyuk Sohn, Ruben Villegas, Gang Pan, Honglak Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 249-258
Object detection systems based on the deep convolutional neural network (CNN) have recently made ground- breaking advances on several object detection benchmarks. While the features learned by these high-capacity neural networks are discriminative for categorization, inaccurate localization is still a major source of error for detection. Building upon high-capacity CNN architectures, we address the localization problem by 1) using a search algorithm based on Bayesian optimization that sequentially proposes candidate regions for an object bounding box, and 2) training the CNN with a structured loss that explicitly penalizes the localization inaccuracy. In experiments, we demonstrated that each of the proposed methods improves the detection performance over the baseline method on PASCAL VOC 2007 and 2012 datasets. Furthermore, two methods are complementary and significantly outperform the previous state-of-the-art when combined.
****

A Metric Parametrization for Trifocal Tensors With Non-Colinear Pinholes
Spyridon Leonardos, Roberto Tron, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 259-267
The trifocal tensor, which describes the relation between projections of points and lines in three views, is a fundamental entity of geometric computer vision. In this work, we investigate a new parametrization of the trifocal tensor for calibrated cameras with non-colinear pinholes obtained from a quotient Riemannian manifold. We incorporate this formulation into state-of-the art methods for optimization on manifolds, and show, through experiments in pose averaging, that it produces a meaningful way to measure distances between trifocal tensors.
****

An Efficient Volumetric Framework for Shape Tracking
Benjamin Allain, Jean-Sebastien Franco, Edmond Boyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 268-276
Recovering 3D shape motion using visual information is an important problem with many applications in computer vision and computer graphics, among other domains

. Most existing approaches rely on surface-based strategies, where surface models are fit to visual surface observations. While numerically plausible, this paradigm ignores the fact that the observed surfaces often delimit volumetric shapes, for which deformations are constrained by the volume inside the shape. Consequently, surface-based strategies can fail when the observations define several feasible surfaces, whereas volumetric considerations are more restrictive with respect to the admissible solutions. In this work, we investigate a novel volumetric shape parametrization to track shapes over temporal sequences. In constrast to Eulerian grid discretizations of the observation space, such as voxels, we consider general shape tesselations yielding more convenient cell decompositions, in particular the Centroidal Voronoi Tesselation. With this shape representation, we devise a tracking method that exploits volumetric information, both for the data term evaluating observation conformity, and for expressing deformation constraints that enforce prior assumptions on motion. Experiments on several datasets demonstrate similar or improved precisions over state-of-the-art methods, as well as improved robustness, a critical issue when tracking sequentially over time frames.

****

Structured Sparse Subspace Clustering: A Unified Optimization Framework
Chun-Guang Li, Rene Vidal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 277-286

Subspace clustering refers to the problem of segmenting data drawn from a union of subspaces. State of the art approaches for solving this problem follow a two-stage approach. In the first step, an affinity matrix is learned from the data using sparse or low-rank minimization techniques. In the second step, the segmentation is found by applying spectral clustering to this affinity. While this approach has led to state of the art results in many applications, it is sub-optimal because it does not exploit the fact that the affinity and the segmentation depend on each other. In this paper, we propose a unified optimization framework for learning both the affinity and the segmentation. Our framework is based on expressing each data point as a structured sparse linear combination of all other data points, where the structure is induced by a norm that depends on the unknown segmentation. We show that both the segmentation and the structured sparse representation can be found via a combination of an alternating direction method of multipliers with spectral clustering. Experiments on a synthetic data, the Hopkins 155 motion segmentation database, and the Extended Yale B data set demonstrate the effectiveness of our approach.

****

Delving Into Egocentric Actions
Yin Li, Zhefan Ye, James M. Rehg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 287-295

We address the challenging problem of recognizing the camera wearer's actions from videos captured by an egocentric camera. Egocentric videos encode a rich set of signals regarding the camera wearer, including head movement, hand pose and gaze information. We propose to utilize these mid-level egocentric cues for egocentric action recognition. We present a novel set of egocentric features and show how they can be combined with motion and object features. The result is a compact representation with superior performance. In addition, we provide the first systematic evaluation of motion, object and egocentric cues in egocentric action recognition. Our benchmark leads to several surprising findings. These findings uncover the best practices for egocentric actions, with a significant performance boost over all previous state-of-the-art methods on three publicly available datasets.

****

Latent Trees for Estimating Intensity of Facial Action Units
Sebastian Kaltwang, Sinisa Todorovic, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 296-304

This paper is about estimating intensity levels of Facial Action Units (FAUs) in videos as an important and challenging step toward interpreting facial expressions. To address uncertainty in detections of facial landmark points, used as out

input features, we formulate a new generative framework comprised of a graphical model, inference, and algorithms for learning both model parameters and structure. Our model is a latent tree (LT) that represents input features of facial landmark points and FAU intensities as leaf nodes, and encodes their higher-order dependencies with latent nodes at tree levels closer to the root. No other restrictions are placed on the model structure beyond that it is a tree. We specify a new algorithm for efficient learning of model structure that iteratively builds LT by adding either new edge or new hidden node to LT, whichever of these two graph-edit operations gives the highest increase of the joint likelihood. Our structure learning efficiently computes likelihood increase and selects an optimal graph revision without considering all possible structural changes. For FAU intensity estimation, we derive closed-form expressions of posterior marginals of all variables in LT, and specify an efficient inference of in two passes -- bottom-up and top-down. Our evaluation on the benchmark DISFA and ShoulderPain datasets, in subject-independent setting, demonstrate that we outperform the state of the art, even in the presence of significant noise in locations of facial landmark points. We demonstrate our correct learning of model structure by probabilistically sampling facial landmark points, conditioned on a given FAU intensity, and thus generating plausible facial expressions.
****

Robust Regression on Image Manifolds for Ordered Label Denoising
Hui Wu, Richard Souvenir; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 305-313
In this paper, we present a computationally efficient and non-parametric method for robust regression on manifolds. We apply our algorithm to the problem of correcting mislabeled examples from image collections with ordered (e.g., real-valued, ordinal) labels. Compared to related methods for robust regression, our method achieves superior denoising accuracy on a variety of data sets, with label corruption levels as high as 80%. For a diverse set of widely-used, large-scale, publicly-available data sets, our approach results in image labels that more accurately describe the associated images.
****

Privacy Preserving Optics for Miniature Vision Sensors
Francesco Pittaluga, Sanjeev J. Koppal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 314-324
The next wave of micro and nano devices will create a world with trillions of small networked cameras. This will lead to increased concerns about privacy and security. Most privacy preserving algorithms for computer vision are applied after image/video data has been captured. We propose to use privacy preserving optics that filter or block sensitive information directly from the incident light-field before sensor measurements are made, adding a new layer of privacy. In addition to balancing the privacy and utility of the captured data, we address trade-offs unique to miniature vision sensors, such as achieving high-quality field-of-view and resolution within the constraints of mass and volume. Our privacy preserving optics enable applications such as depth sensing, full-body motion tracking, people counting, blob detection and privacy preserving face recognition. While we demonstrate applications on macro-scale devices (smartphones, webcams, etc.) our theory has impact for smaller devices.
****

Deep Transfer Metric Learning
Junlin Hu, Jiwen Lu, Yap-Peng Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 325-333
Conventional metric learning methods usually assume that the training and test samples are captured in similar scenarios so that their distributions are assumed to be the same. This assumption doesn't hold in many real visual recognition applications, especially when samples are captured across different datasets. In this paper, we propose a new deep transfer metric learning (DTML) method to learn a set of hierarchical nonlinear transformations for cross-domain visual recognition by transferring discriminative knowledge from the labeled source domain to the unlabeled target domain. Specifically, our DTML learns a deep metric network

by maximizing the inter-class variations and minimizing the intra-class variations, and minimizing the distribution divergence between the source domain and the target domain at the top layer of the network. To better exploit the discriminative information from the source domain, we further develop a deeply supervised transfer metric learning (DSTML) method by including an additional objective on DTML where the output of both the hidden layers and the top layer are optimized jointly. Experimental results on cross-dataset face verification and person re-identification validate the effectiveness of the proposed methods.
****

Small-Variance Nonparametric Clustering on the Hypersphere
Julian Straub, Trevor Campbell, Jonathan P. How, John W. Fisher III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 334-342
Structural regularities in man-made environments reflect in the distribution of their surface normals. Describing these surface normal distributions is important in many computer vision applications, such as scene understanding, plane segmentation, and regularization of 3D reconstructions. Based on the small-variance limit of Bayesian nonparametric von-Mises-Fisher (vMF) mixture distributions, we propose two new flexible and efficient k-means-like clustering algorithms for directional data such as surface normals. The first, DP-vMF-means, is a batch clustering algorithm derived from the Dirichlet process (DP) vMF mixture. Recognizing the sequential nature of data collection in many applications, we extend this algorithm to DDP-vMF-means, which infers temporally evolving cluster structure from streaming data. Both algorithms naturally respect the geometry of directional data, which lies on the unit sphere. We demonstrate their performance on synthetic directional data and real 3D surface normals from RGB-D sensors. While our experiments focus on 3D data, both algorithms generalize to high dimensional directional data such as protein backbone configurations and semantic word vectors.
****

DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time
Richard A. Newcombe, Dieter Fox, Steven M. Seitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 343-352
We present the first dense SLAM system capable of reconstructing non-rigidly deforming scenes in real-time, by fusing together RGBD scans captured from commodity sensors. Our DynamicFusion approach reconstructs scene geometry whilst simultaneously estimating a dense volumetric 6D motion field that warps the estimated geometry into a live frame. Like KinectFusion, our system produces increasingly denoised, detailed, and complete reconstructions as more measurements are fused, and displays the updated model in real time. Because we do not require a template or other prior scene model, the approach is applicable to a wide range of moving objects and scenes.
****

Reliable Patch Trackers: Robust Visual Tracking by Exploiting Reliable Patches
Yang Li, Jianke Zhu, Steven C.H. Hoi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 353-361
Most modern trackers typically employ a bounding box given in the first frame to track visual objects, where their tracking results are often sensitive to the initialization. In this paper, we propose a new tracking method, Reliable Patch Trackers (RPT), which attempts to identify and exploit the reliable patches that can be tracked effectively through the whole tracking process. Specifically, we present a tracking reliability metric to measure how reliably a patch can be tracked, where a probability model is proposed to estimate the distribution of reliable patches under a sequential Monte Carlo framework. As the reliable patches distributed over the image, we exploit the motion trajectories to distinguish them from the background. Therefore, the visual object can be defined as the clustering of homo-trajectory patches, where a Hough voting-like scheme is employed to estimate the target state. Encouraging experimental results on a large set of sequences showed that the proposed approach is very effective and in comparison to the state-of-the-art trackers. The full source code of our implementation will be publicly available.

****
Predicting Eye Fixations Using Convolutional Neural Networks
Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, Tianming Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 362-370
It is believed that eye movements in free-viewing of natural scenes are directed by both bottom-up visual saliency and top-down visual factors. In this paper, we propose a novel computational framework to simultaneously learn these two types of visual features from raw image data using a multiresolution convolutional neural network (Mr-CNN) for predicting eye fixations. The Mr-CNN is directly trained from image regions centered on fixation and non-fixation locations over multiple resolutions, using raw image pixels as inputs and eye fixation attributes as labels. Diverse top-down visual features can be learned in higher layers. Meanwhile bottom-up visual saliency can also be inferred via combining information over multiple resolutions. Finally, optimal integration of bottom-up and top-down cues can be learned in the last logistic regression layer to predict eye fixations. The proposed approach achieves state-of-the-art results over four publically available benchmark datasets, demonstrating the superiority of our work.
****
Kernel Fusion for Better Image Deblurring
Long Mai, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 371-380
Kernel estimation for image deblurring is a challenging task and a large number of algorithms have been developed. Our hypothesis is that while individual kernels estimated using different methods alone are sometimes inadequate, they often complement each other. This paper addresses the problem of fusing multiple kernels estimated using different methods into a more accurate one that can better support image deblurring than each individual kernel. In this paper, we develop a data-driven approach to kernel fusion that learns how each kernel contributes to the final kernel and how they interact with each other. We discuss various kernel fusion models and find that kernel fusion using Gaussian Conditional Random Fields performs best. This Gaussian Conditional Random Fields-based kernel fusion method not only models how individual kernels are fused at each kernel element but also the interaction of kernel fusion among multiple kernel elements. Our experiments show that our method can significantly improve image deblurring by combining kernels from multiple methods into a better one.
****
Direction Matters: Depth Estimation With a Surface Normal Classifier
Christian Hane, Lubor Ladicky, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 381-389
In this work we make use of recent advances in data driven classification to improve standard approaches for binocular stereo matching and single view depth estimation. Surface normal direction estimation has become feasible and shown to work reliably on state of the art benchmark datasets. Information about the surface orientation contributes crucial information about the scene geometry in cases where standard approaches struggle. We describe, how the responses of such a classifier can be included in global stereo matching approaches. One of the strengths of our approach is, that we can use the classifier responses for a whole set of directions and let the final optimization decide about the surface orientation. This is important in cases where based on the classifier, multiple different surface orientations seem likely. We evaluate our method on two challenging real-world datasets for the two proposed applications. For the binocular stereo matching we use road scene imagery taken from a car and for the single view depth estimation we use images taken in indoor environments.
****
Modeling Local and Global Deformations in Deep Learning: Epitomic Convolution, Multiple Instance Learning, and Sliding Window Detection
George Papandreou, Iasonas Kokkinos, Pierre-Andre Savalle; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 390-399

Deep Convolutional Neural Networks (DCNNs) achieve invariance to domain transformations (deformations) by using multiple 'max-pooling' (MP) layers. In this work we show that alternative methods of modeling deformations can improve the accuracy and efficiency of DCNNs. First, we introduce epitomic convolution as an alternative to the common convolution-MP cascade of DCNNs, that comes with the same computational cost but favorable learning properties. Second, we introduce a Multiple Instance Learning algorithm to accommodate global translation and scaling in image classification, yielding an efficient algorithm that trains and tests a DCNN in a consistent manner. Third we develop a DCNN sliding window detector that explicitly, but efficiently, searches over the object's position, scale, and aspect ratio. We provide competitive image classification and localization results on the ImageNet dataset and object detection results on Pascal VOC2007.
****

Grasp Type Revisited: A Modern Perspective on a Classical Feature for Vision
Yezhou Yang, Cornelia Fermuller, Yi Li, Yiannis Aloimonos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 400-408
The grasp type provides crucial information about human action. However, recognizing the grasp type in unconstrained scenes is challenging because of the large variations in appearance, occlusions and geometric distortions. In this paper, first we present a convolutional neural network to classify functional hand grasp types. Experiments on a public static scene hand data set validate good performance of the presented method. Then we present two applications utilizing grasp type classification: (a) inference of human action intention and (b) fine level manipulation action segmentation. Experiments on both tasks demonstrate the usefulness of grasp type as a cognitive feature for computer vision. This study shows that the grasp type is a powerful symbolic representation for action understanding, and thus opens new avenues for future research.
****

Learning Hypergraph-Regularized Attribute Predictors
Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, Dan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 409-417
We present a novel attribute learning framework named Hypergraph-based Attribute Predictor (HAP). In HAP, a hypergraph is leveraged to depict the attribute relations in the data. Then the attribute prediction problem is casted as a regularized hypergraph cut problem, in which a collection of attribute projections is jointly learnt from the feature space to a hypergraph embedding space aligned with the attributes. The learned projections directly act as attribute classifiers (linear and kernelized). This formulation leads to a very efficient approach. By considering our model as a multi-graph cut task, our framework can flexibly incorporate other available information, in particular class label. We apply our approach to attribute prediction, Zero-shot and N-shot learning tasks. The results on AWA, USAA and CUB databases demonstrate the value of our methods in comparison with the state-of-the-art approaches.
****

A Coarse-to-Fine Model for 3D Pose Estimation and Sub-Category Recognition
Roozbeh Mottaghi, Yu Xiang, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 418-426
Despite the fact that object detection, 3D pose estimation, and sub-category recognition are highly correlated tasks, they are usually addressed independently from each other because of the huge space of parameters. To jointly model all of these tasks, we propose a coarse-to-fine hierarchical representation, where each level of the hierarchy represents objects at a different level of granularity. The hierarchical representation prevents performance loss, which is often caused by the increase in the number of parameters (as we consider more tasks to model), and the joint modeling enables resolving ambiguities that exist in independent modeling of these tasks. We augment PASCAL 3D+ dataset with annotations for these tasks and show that our hierarchical model is effective in joint modeling of object detection, 3D pose estimation, and sub-category recognition.

****
Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images
Anh Nguyen, Jason Yosinski, Jeff Clune; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 427-436
Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which we call "fooling images" (more generally, fooling examples). Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.
****
Deformable Part Models are Convolutional Neural Networks
Ross Girshick, Forrest Iandola, Trevor Darrell, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 437-446
Deformable part models (DPMs) and convolutional neural networks (CNNs) are two widely used tools for visual recognition. They are typically viewed as distinct approaches: DPMs are graphical models (Markov random fields), while CNNs are "black-box" non-linear classifiers. In this paper, we show that a DPM can be formulated as a CNN, thus providing a synthesis of the two ideas. Our construction involves unrolling the DPM inference algorithm and mapping each step to an equivalent CNN layer. From this perspective, it is natural to replace the standard image features used in DPMs with a learned feature extractor. We call the resulting model a DeepPyramid DPM and experimentally validate it on PASCAL VOC object detection. We find that DeepPyramid DPMs significantly outperform DPMs based on histograms of oriented gradients features (HOG) and slightly outperforms a comparable version of the recently introduced R-CNN detection system, while running significantly faster.
****
Hypercolumns for Object Segmentation and Fine-Grained Localization
Bharath Hariharan, Pablo Arbelaez, Ross Girshick, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 447-456
Recognition algorithms based on convolutional networks (CNNs) typically use the output of the last layer as feature representation. However, the information in this layer may be too coarse to allow precise localization. On the contrary, earlier layers may be precise in localization but will not capture semantics. To get the best of both worlds, we define the hypercolumn at a pixel as the vector of activations of all CNN units above that pixel. Using hypercolumns as pixel descriptors, we show results on three fine-grained localization tasks: simultaneous detection and segmentation[20], where we improve state-of-the-art from 49.7 mean AP^r[20] to 59.0, keypoint localization, where we get a 3.3 point boost over [19] and part labeling, where we show a 6.6 point gain over a strong baseline.
****
Mapping Visual Features to Semantic Profiles for Retrieval in Medical Imaging
Johannes Hofmanninger, Georg Langs; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 457-465

Content based image retrieval is highly relevant in medical imaging, since it ma
kes vast amounts of imaging data accessible for comparison during diagnosis. Fin
ding image similarity measures that reflect diagnostically relevant relationship
s is challenging, since the overall appearance variability is high compared to o
ften subtle signatures of diseases. To learn models that capture the relationshi
p between semantic clinical information and image elements at scale, we have to
rely on data generated during clinical routine (images and radiology reports), s
ince expert annotation is prohibitively costly. Here we show that re-mapping vis
ual features extracted from medical imaging data based on weak labels that can b
e found in corresponding radiology reports creates descriptions of local image c
ontent capturing clinically relevant information. We show that these semantic pr
ofiles enable higher recall and precision during retrieval compared to visual fe
atures, and that we can even map semantic terms describing clinical findings fro
m radiology reports to localized image volume areas.
****

Event-Driven Stereo Matching for Real-Time 3D Panoramic Vision
Stephan Schraml, Ahmed Nabil Belbachir, Horst Bischof; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 466-474
This paper presents a stereo matching approach for a novel multi-perspective pan
oramic stereo vision system, making use of asynchronous and non-simultaneous ste
reo imaging towards real-time 3D 360deg vision. The method is designed for event
s representing the scenes visual contrast as a sparse visual code allowing the s
tereo reconstruction of high resolution panoramic views. We propose a novel cost
measure for the stereo matching, which makes use of a similarity measure based
on event distributions. Thus, the robustness to variations in event occurrences
was increased. An evaluation of the proposed stereo method is presented using di
stance estimation of panoramic stereo views and ground truth data. Furthermore,
our approach is compared to standard stereo methods applied on event-data. Resul
ts show that we obtain 3D reconstructions of 1024 x 3600 round views and outperf
orm depth reconstruction accuracy of state-of-the-art methods on event data.
****

Graph-Based Simplex Method for Pairwise Energy Minimization With Binary Variable
s
Daniel Prusa; Proceedings of the IEEE Conference on Computer Vision and Pattern
Recognition (CVPR), 2015, pp. 475-483
We show how the simplex algorithm can be tailored to the linear programming rela
xation of pairwise energy minimization with binary variables. A special structur
e formed by basic and nonbasic variables in each stage of the algorithm is ident
ified and utilized to perform the whole iterative process combinatorially over t
he input energy minimization graph rather than algebraically over the simplex ta
bleau. This leads to a new efficient solver. We demonstrate that for some comput
er vision instances it performs even better than methods reducing binary energy
minimization to finding maximum flow in a network.
****

Image Denoising via Adaptive Soft-Thresholding Based on Non-Local Samples
Hangfan Liu, Ruiqin Xiong, Jian Zhang, Wen Gao; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 484-492
This paper proposes a new image denoising approach using adaptive signal modelin
g and adaptive soft-thresholding. It improves the image quality by regularizing
all the patches in image based on distribution modeling in transform domain. Ins
tead of using a global model for all patches, it employs content adaptive models
to address the non-stationarity of image signals. The distribution model of eac
h patch is estimated individually and can vary for different transform bands and
for different patch locations. In particular, we allow the distribution model f
or each individual patch to have non-zero expectation. To estimate the expectati
on and variance parameters for the transform bands of a particular patch, we exp
loit the non-local correlation of image and collect a set of similar patches as
data samples to form the distribution. Irrelevant patches are excluded so that t
his non-local based modeling is more accurate than global modeling. Adaptive sof
t-thresholding is employed since we observed that the distribution of non-local

samples can be approximated by Laplacian distribution. Experimental results show that the proposed scheme outperforms the state-of-the-art denoising methods such as BM3D and CSR in both the PSNR and the perceptual quality.
****

3D Scanning Deformable Objects With a Single RGBD Sensor
Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, Shahram Izadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 493-501
We present a 3D scanning system for deformable objects that uses only a single Kinect sensor. Our work allows considerable amount of nonrigid deformations during scanning, and achieves high quality results without heavily constraining user or camera motion. We do not rely on any prior shape knowledge, enabling general object scanning with freeform deformations. To deal with the drift problem when nonrigidly aligning the input sequence, we automatically detect loop closures, distribute the alignment error over the loop, and finally use a bundle adjustment algorithm to optimize for the latent 3D shape and nonrigid deformation parameters simultaneously. We demonstrate high quality scanning results in some challenging sequences, comparing with state of art nonrigid techniques, as well as ground truth data.
****

Nested Motion Descriptors
Jeffrey Byrne; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 502-510
A nested motion descriptor is a spatiotemporal representation of motion that is invariant to global camera translation, without requiring an explicit estimate of optical flow or camera stabilization. This descriptor is a natural spatiotemporal extension of the nested shape descriptor to the representation of motion. We demonstrate that the quadrature steerable pyramid can be used to pool phase, and that pooling phase rather than magnitude provides an estimate of camera motion. This motion can be removed using the log-spiral normalization as introduced in the nested shape descriptor. Furthermore, this structure enables an elegant visualization of salient motion using the reconstruction properties of the steerable pyramid. We compare our descriptor to local motion descriptors, HOG-3D and HOG-HOF, and show improvements on three activity recognition datasets.
****

Efficient Minimal-Surface Regularization of Perspective Depth Maps in Variational Stereo
Gottfried Graber, Jonathan Balzer, Stefano Soatto, Thomas Pock; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 511-520
We propose a method for dense three-dimensional surface reconstruction that leverages the strengths of shape-based approaches, by imposing regularization that respects the geometry of the surface, and the strength of depth-map-based stereo, by avoiding costly computation of surface topology. The result is a near real-time variational reconstruction algorithm free of the staircasing artifacts that affect depth-map and plane-sweeping approaches. This is made possible by exploiting the gauge ambiguity to design a novel representation of the regularizer that is linear in the parameters and hence amenable to be optimized with state-of-the-art primal-dual numerical schemes.
****

Maximum Persistency via Iterative Relaxed Inference With Graphical Models
Alexander Shekhovtsov, Paul Swoboda, Bogdan Savchynskyy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 521-529
We consider MAP-inference for graphical models and propose a novel efficient algorithm for finding persistent labels. Our algorithm marks each label in each node of the considered graphical model either as (i) optimal, meaning that it belongs to all optimal solutions of the inference problem; (ii) non-optimal if it provably does not belong to any solution; or (iii) undefined, which means our algorithm can not make a decision regarding the label. Moreover, we prove optimality of our approach, that it delivers in a certain sense the largest total number o

f labels marked as optimal or non-optimal. We demonstrate superiority of our app
roach on problems from machine learning and computer vision benchmarks.
****

Deep Hierarchical Parsing for Semantic Segmentation
Abhishek Sharma, Oncel Tuzel, David W. Jacobs; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 530-538
This paper proposes a learning-based approach to scene parsing inspired by the d
eep Recursive Context Propagation Network (RCPN). RCPN is a deep feed-forward ne
ural network that utilizes the contextual information from the entire image, thr
ough bottom-up followed by top-down context propagation via random binary parse
trees. This improves the feature representation of every super-pixel in the imag
e for better classification into semantic categories. We analyze RCPN and propos
e two novel contributions to further improve the model. We first analyze the lea
rning of RCPN parameters and discover the presence of bypass error paths in the
computation graph of RCPN that can hinder contextual propagation. We propose to
tackle this problem by including the classification loss of the internal nodes o
f the random parse trees in the original RCPN loss function. Secondly, we use an
 MRF on the parse tree nodes to model the hierarchical dependency present in the
 output. Both modifications provide performance boosts over the original RCPN an
d the new system achieves state-of-the-art performance on Stanford Background, S
IFT-Flow and Daimler urban datasets.
****

Designing Deep Networks for Surface Normal Estimation
Xiaolong Wang, David Fouhey, Abhinav Gupta; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2015, pp. 539-547
In the past few years, convolutional neural nets (CNN) have shown incredible pro
mise for learning visual representations. In this paper, we use CNNs for the tas
k of predicting surface normals from a single image. But what is the right archi
tecture? We propose to build upon the decades of hard work in 3D scene understan
ding to design a new CNN architecture for the task of surface normal estimation.
 We show that incorporating several constraints (man-made, Manhattan world) and
meaningful intermediate representations (room layout, edge labels) in the archit
ecture leads to state of the art performance on surface normal estimation. We al
so show that our network is quite robust and show state of the art results on ot
her datasets as well without any fine-tuning.
****

Layered RGBD Scene Flow Estimation
Deqing Sun, Erik B. Sudderth, Hanspeter Pfister; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 548-556
As consumer depth sensors become widely available, estimating scene flow from RG
BD sequences has received increasing attention. Although the depth information a
llows the recovery of 3D motion from a single view, it poses new challenges. In
particular, depth boundaries are not well-aligned with RGB image edges and there
fore not reliable cues to localize 2D motion boundaries. In addition, methods t
hat extend the 2D optical flow formulation to 3D still produce large errors in o
cclusion regions. To better use depth for occlusion reasoning, we propose a laye
red RGBD scene flow method that jointly solves for the scene segmentation and th
e motion. Our key observation is that the noisy depth is sufficient to decide th
e depth ordering of layers, thereby avoiding a computational bottleneck for RGB
layered methods. Furthermore, the depth enables us to estimate a per-layer 3D r
igid motion to constrain the motion of each layer. Experimental results on both
the Middlebury and real-world sequences demonstrate the effectiveness of the lay
ered approach for RGBD scene flow estimation.
****

Hashing With Binary Autoencoders
Miguel A. Carreira-Perpinan, Ramin Raziperchikolaei; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 557-566
An attractive approach for fast search in image databases is binary hashing, whe
re each high-dimensional, real-valued image is mapped onto a low-dimensional, bi
nary vector and the search is done in this binary space. Finding the optimal has

h function is difficult because it involves binary constraints, and most approac hes approximate the optimization by relaxing the constraints and then binarizing the result. Here, we focus on the binary autoencoder model, which seeks to reco nstruct an image from the binary code produced by the hash function. We show tha t the optimization can be simplified with the method of auxiliary coordinates. T his reformulates the optimization as alternating two easier steps: one that lear ns the encoder and decoder separately, and one that optimizes the code for each image. Image retrieval experiments show the resulting hash function outperforms or is competitive with state-of-the-art methods for binary hashing.
****

SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite
Shuran Song, Samuel P. Lichtenberg, Jianxiong Xiao; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 567-576
Although RGB-D sensors have enabled major breakthroughs for several vision tasks , such as 3D reconstruction, we have not attained the same level of success in h igh-level scene understanding. Perhaps one of the main reasons is the lack of a large-scale benchmark with 3D annotations and 3D evaluation metrics. In this pap er, we introduce an RGB-D benchmark suite for the goal of advancing the state-of -the-arts in all major scene understanding tasks. Our dataset is captured by fou r different sensors and contains 10,335 RGB-D images, at a similar scale as PASC AL VOC. The whole dataset is densely annotated and includes 146,617 2D polygons and 64,595 3D bounding boxes with accurate object orientations, as well as a 3D room layout and scene category for each image. This dataset enables us to train data-hungry algorithms for scene-understanding tasks, evaluate them using meanin gful 3D metrics, avoid overfitting to a small testing set, and study cross-senso r bias.
****

Collaborative Feature Learning From Social Media
Chen Fang, Hailin Jin, Jianchao Yang, Zhe Lin; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 577-585
Image feature representation plays an essential role in image recognition and re lated tasks. The current state-of-the-art feature learning paradigm is supervise d learning from labeled data. However, this paradigm requires large-scale catego ry labels, which limits its applicability to domains where labels are hard to ob tain. In this paper, we propose a new data-driven feature learning paradigm whic h does not rely on category labels. Instead, we learn from user behavior data co llected on social media. Concretely, we use the image relationship discovered in the latent space from the user behavior data to guide the image feature learnin g. We collect a large-scale image and user behavior dataset from Behance.net. Th e dataset consists of 1.9 million images and over 300 million view records from 1.9 million users. We validate our feature learning paradigm on this dataset and find that the learned feature significantly outperforms the state-of-the-art im age features in learning better image similarities. We also show that the learne d feature performs competitively on various recognition benchmarks.
****

Diversity-Induced Multi-View Subspace Clustering
Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, Hua Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 586 -594
In this paper, we focus on how to boost the multi-view clustering by exploring t he complementary information among multi-view features. A multi-view clustering framework, called Diversity-induced Multi-view Subspace Clustering (DiMSC), is p roposed for this task. In our method, we extend the existing subspace clustering into the multi-view domain, and utilize the Hilbert Schmidt Independence Criter ion (HSIC) as a diversity term to explore the complementarity of multi-view repr esentations, which could be solved efficiently by using the alternating minimizi ng optimization. Compared to other multi-view clustering methods, the enhanced c omplementarity reduces the redundancy between the multi-view features, and impro ves the accuracy of the clustering results. Experiments on both image and video face clustering well demonstrate that the proposed method outperforms the state-

of-the-art methods.
****

Building a Bird Recognition App and Large Scale Dataset With Citizen Scientists: The Fine Print in Fine-Grained Dataset Collection
Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 595-604
We introduce tools and methodologies to collect high quality, large scale fine-grained computer vision datasets using citizen scientists -- crowd annotators who are passionate and knowledgeable about specific domains such as birds or airplanes. We worked with citizen scientists and domain experts to collect NABirds, a new high quality dataset containing 48,562 images of North American birds with 555 categories, part annotations and bounding boxes. We find that citizen scientists are significantly more accurate than Mechanical Turkers at zero cost. We worked with bird experts to measure the quality of popular datasets like CUB-200-2011 and ImageNet and found class label error rates of at least 4%. Nevertheless, we found that learning algorithms are surprisingly robust to annotation errors and this level of training data corruption can lead to an acceptably small increase in test error if the training set has sufficient size. At the same time, we found that an expert-curated high quality test set like NABirds is necessary to accurately measure the performance of fine-grained computer vision systems. We used NABirds to train a publicly available bird recognition service deployed on the web site of the Cornell Lab of Ornithology.
****

Early Burst Detection for Memory-Efficient Image Retrieval
Miaojing Shi, Yannis Avrithis, Herve Jegou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 605-613
Recent works show that image comparison based on local descriptors is corrupted by visual bursts, which tend to dominate the image similarity. The existing strategies, like power-law normalization, improve the results by discounting the contribution of visual bursts to the image similarity. In this paper, we propose to explicitly detect the visual bursts in an image at an early stage. We compare several detection strategies jointly taking into account feature similarity and geometrical quantities. The bursty groups are merged into meta-features, which are used as input to state-of-the-art image search systems such as VLAD or the selective match kernel. Then, we show the interest of using this strategy in an asymmetrical manner, with only the database features being aggregated but not those of the query. Extensive experiments performed on public benchmarks for visual retrieval show the benefits of our method, which achieves performance on par with the state of the art but with a significantly reduced complexity, thanks to the lower number of features fed to the indexing system.
****

Indoor Scene Structure Analysis for Single Image Depth Estimation
Wei Zhuo, Mathieu Salzmann, Xuming He, Miaomiao Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 614-622
We tackle the problem of single image depth estimation, which, without additional knowledge, suffers from many ambiguities. Unlike previous approaches that only reason locally, we propose to exploit the global structure of the scene to estimate its depth. To this end, we introduce a hierarchical representation of the scene, which models local depth jointly with mid-level and global scene structures. We formulate single image depth estimation as inference in a graphical model whose edges let us encode the interactions within and across the different layers of our hierarchy. Our method therefore still produces detailed depth estimates, but also leverages higher-level information about the scene. We demonstrate the benefits of our approach over local depth estimation methods on standard indoor datasets.
****

Light Field Layer Matting
Juliet Fiss, Brian Curless, Rick Szeliski; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 623-631

In this paper, we use matting to separate foreground layers from light fields ca ptured with a plenoptic camera. We represent the input 4D light field as a 4D ba ckground light field, plus a 2D spatially varying foreground color layer with al pha. Our method can be used to both pull a fore- ground matte and estimate an oc cluded background light field. Our method assumes that the foreground layer is t hin and fronto-parallel, and is composed of a limited set of col- ors that are d istinct from the background layer colors. Our method works well for thin, transl ucent, and blurred foreground occluders. Our representation can be used to rende r the light field from novel views, handling disocclusions while avoiding common artifacts.
****

Depth Camera Tracking With Contour Cues
Qian-Yi Zhou, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2015, pp. 632-638
We present an approach for tracking camera pose in real time given a stream of d epth images. Existing algorithms are prone to drift in the presence of smooth su rfaces that destabilize geometric alignment. We show that useful contour cues ca n be extracted from noisy and incomplete depth input. These cues are used to est ablish correspondence constraints that carry information about scene geometry an d constrain pose estimation. Despite ambiguities in the input, the presented con tour constraints reliably improve tracking accuracy. Results on benchmark sequen ces and on additional challenging examples demonstrate the utility of contour cu es for real-time camera pose estimation.
****

Radial Distortion Homography
Zuzana Kukelova, Jan Heller, Martin Bujnak, Tomas Pajdla; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 639-64 7
The importance of precise homography estimation is often underestimated even tho ugh it plays a crucial role in various vision applications such as plane or plan arity detection, scene degeneracy tests, camera motion classification, image sti tching, and many more. Ignoring the radial distortion component in homography e stimation---even for classical perspective cameras---may lead to significant err ors or totally wrong estimates. In this paper, we fill the gap among the homogra phy estimation methods by presenting two algorithms for estimating homography be tween two cameras with different radial distortions. Both algorithms can handle planar scenes as well as scenes where the relative motion between the cameras is a pure rotation. The first algorithm uses the minimal number of five image poin t correspondences and solves a nonlinear system of polynomial equations using th e Groebner basis method. The second algorithm uses a non-minimal number of six i mage point correspondences and leads to a simple system of two quadratic equatio ns in two unknowns and one system of six linear equations. The proposed algorith ms are fast, stable, and can be efficiently used inside a RANSAC loop.
****

Efficient Object Localization Using Convolutional Networks
Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christoph Bregler; Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2015, pp. 648-656
Recent state-of-the-art performance on human-body pose estimation has been achie ved with Deep Convolutional Networks (ConvNets). Traditional ConvNet architectur es include pooling and sub-sampling layers which reduce computational requiremen ts, introduce invariance and prevent over-training. These benefits of pooling co me at the cost of reduced localization accuracy. We introduce a novel architectu re which includes an efficient `position refinement' model that is trained to es timate the joint offset location within a small region of the image. This refine ment model is jointly trained in cascade with a state-of-the-art ConvNet model t o achieve improved accuracy in human joint location estimation. We show that the variance of our detector approaches the variance of human annotations on the FL IC dataset and outperforms all existing approaches on the MPII-human-pose datase t.

****
Just Noticeable Defocus Blur Detection and Estimation
Jianping Shi, Li Xu, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 657-665
We tackle a fundamental problem to detect and estimate just noticeable blur (JNB) caused by defocus that spans a small number of pixels in images. This type of blur is common during photo taking. Although it is not strong, the slight edge blurriness contains informative clues related to depth. We found existing blur descriptors based on local information cannot distinguish this type of small blur reliably from unblurred structures. We propose a simple yet effective blur feature via sparse representation and image decomposition. It directly establishes correspondence between sparse edge representation and blur strength estimation. Extensive experiments manifest the generality and robustness of this feature.
****
How Do We Use Our Hands? Discovering a Diverse Set of Common Grasps
De-An Huang, Minghuang Ma, Wei-Chiu Ma, Kris M. Kitani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 666-675
Our aim is to show how state-of-the-art computer vision techniques can be used to advance prehensile analysis (i.e., understanding the functionality of human hands). Prehensile analysis is a broad field of multi-disciplinary interest, where researchers painstakingly manually analyze hours of hand-object interaction videos to understand the mechanics of hand manipulation. In this work, we present promising empirical results indicating that wearable cameras and unsupervised clustering techniques can be used to automatically discover common modes of human hand use. In particular, we use a first-person point-of-view camera to record common manipulation tasks and leverage its strengths for reliably observing human hand use. To learn a diverse set of hand-object interactions, we propose a fast online clustering algorithm based on the Determinantal Point Process (DPP). Furthermore, we develop a hierarchical extension to the DPP clustering algorithm and show that it can be used to discover appearance-based grasp taxonomies. Using a purely data-driven approach, our proposed algorithm is able to obtain hand grasp taxonomies that roughly correspond to the classic Cutkosky grasp taxonomy. We validate our approach on over 10 hours of first-person point-of-view videos in both choreographed and real-life scenarios.
****
Rotating Your Face Using Multi-Task Deep Neural Network
Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 676-684
Face recognition under viewpoint and illumination changes is a difficult problem, so many researchers have tried to solve this problem by producing the pose- and illumination- invariant feature. Zhu et al. [26] changed all arbitrary pose and illumination images to the frontal view image to use for the invariant feature. In this scheme, preserving identity while rotating pose image is a crucial issue. This paper proposes a new deep architecture based on a novel type of multitask learning, which can achieve superior performance in rotating to a target-pose face image from an arbitrary pose and illumination image while preserving identity. The target pose can be controlled by the user's intention. This novel type of multi-task model significantly improves identity preservation over the single task model. By using all the synthesized controlled pose images, called Controlled Pose Image (CPI), for the pose- illumination- invariant feature and voting among the multiple face recognition results, we clearly outperform the state-of-the-art algorithms by more than 4~6% on the MultiPIE dataset.
****
Is Object Localization for Free? - Weakly-Supervised Learning With Convolutional Neural Networks
Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 685-694
Successful visual object recognition methods typically rely on training datasets containing lots of richly annotated images. Annotating object bounding boxes is

both expensive and subjective. We describe a weakly supervised convolutional neural network (CNN) for object classification that relies only on image-level labels, yet can learn from cluttered scenes containing multiple objects. We quantify its object classification and object location prediction performance on the Pascal VOC 2012 (20 object classes) and the much larger Microsoft COCO (80 object classes) datasets. We find that the network (i) outputs accurate image-level labels, (ii) predicts approximate locations (but not extents) of objects, and (iii) performs similar or better compared to its fully-supervised counterparts using object bounding box annotation for training.
****

Super-Resolution Person Re-Identification With Semi-Coupled Low-Rank Discriminant Dictionary Learning
Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, Baowen Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 695-704
Person re-identification has been widely studied due to its importance in surveillance and forensics applications. In practice, gallery images are high-resolution (HR) while probe images are usually low-resolution (LR) in the identification scenarios with large variation of illumination, weather or quality of cameras. Person re-identification in this kind of scenarios, which we call super-resolution (SR) person re-identification, has not been well studied. In this paper, we propose a semi-coupled low-rank discriminant dictionary learning (SLD^2L) approach for SR person re-identification. For the given training image set which consists of HR gallery and LR probe images, we aim to convert the features of LR images into discriminating HR features. Specifically, our approach learns a pair of HR and LR dictionaries and a mapping from the features of HR gallery images and LR probe images. To ensure that the converted features using the learned dictionaries and mapping have favorable discriminative capability, we design a discriminant term which requires the converted HR features of LR probe images should be close to the features of HR gallery images from the same person, but far away from the features of HR gallery images from different persons. In addition, we apply low-rank regularization in dictionary learning procedure such that the learned dictionaries can well characterize intrinsic feature space of HR and LR images. Experimental results on public datasets demonstrate the effectiveness of SLD^2L.
****

Notice of Violation of IEEE Publication Principles: Dual Domain Filters Based Texture and Structure Preserving Image Non-Blind Deconvolution
Hang Yang, Ming Zhu, Yan Niu, Yujing Guan, Zhongbo Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 705-713
The following message is relayed from an update made on IEEE Xplore.

Notice of Violation of IEEE Publication Principles

"Dual Domain Filters Based Texture and Structure Preserving Image Non-Blind Deconvolution"
by Hang Yang, Ming Zhu, Yan Niu, Yujing Guan, and Zhongbo Zhang
in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 705-713

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEEâ■■s Publication Principles.

This paper copied portions of text from the papers cited below. The original text was copied without attribution (including appropriate references to the original author(s) and/or paper titles) and without permission.

"Group-Based Sparse Representation for Image Restoration"

by Jian Zhang, Debin Zhao, and Wen Gao
in the IEEE Transactions on Image Processing, Vol 23, No 8, August 2014, pp. 333
6-3351

"Dual-domain Image Denoising"
by Claude Knaus, Matthias Zwicker
in the Proceedings of the IEEE International Conference on Image Processing, (IC
IP), September 2013, pp. 440-444

"A Machine Learning Approach for Non-blind Image Deconvolution"
by Christian Schuler, Harold Christopher Burger, Stefan Harmeling, and Bernhard
Scholkopf
in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni
tion (CVPR), June 2013, pp. 1067-1074

Image deconvolution continues to be an active research topic of recovering a sha
rp image, given a blurry one generated by a convolution. One of the most challen
ging problems in image deconvolution is how to preserve the fine scale texture s
tructures while removing blur and noise. Various methods have been implemented i
n both spatial and transform domains, such as gradient based methods, nonlocal s
elf-similarity methods, sparsity based methods. However, each domain has its ad
vantages and shortcomings, which can be complemented by each other. In this work
 we propose a new approach for efficient image deconvolution based on dual domai
n filters. In the deblurring process, we offer a hybrid method that a novel roll
ing guidance filter is used to ensure proper texture/structure separation, and t
hen in the transform domain, we use the short-time Fourier transform to recover
the textures while removing noise with energy shrinkage. Our hybrid algorithm th
at is surprisingly easy to implement, and experimental results clearly show that
 the proposed algorithm outperforms many state-of-the-art deconvolution algorith
ms in terms of both quantitative measure and visual perception quality.
****

Region-Based Temporally Consistent Video Post-Processing
Xuan Dong, Boyan Bonev, Yu Zhu, Alan L. Yuille; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 714-722
We study the problem of temporally consistent video post-processing. Previous po
st-processing algorithms usually either fail to keep high fidelity or fail to ke
ep temporal consistency of output videos. In this paper, we observe experimental
ly that many image/video enhancement algorithms enforce a spatially consistent p
rior on the enhancement. More precisely, within a local region, the enhancement
is consistent, i.e., pixels with the same RGB values will get the same enhanceme
nt values. Using this prior, we segment each frame into several regions and temp
orally-spatially adjust the enhancement of regions of different frames, taking i
nto account fidelity, temporal consistency and spatial consistency. User study,
objective measurement and visual quality comparisons are conducted. The experime
ntal results demonstrate that our output videos can keep high fidelity and tempo
ral consistency at the same time.
****

Global Refinement of Random Forest
Shaoqing Ren, Xudong Cao, Yichen Wei, Jian Sun; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 723-730
Random forest is well known as one of the best learning methods. In spite of its
 great success, it also has certain drawbacks: the heuristic learning rule does
not effectively minimize the global training loss; the model size is usually too
 large for many real applications. To address the issues, we propose two techniq
ues, global refinement and global pruning, to improve a pre-trained random fores
t. The proposed global refinement jointly relearns the leaf nodes of all trees u
nder a global objective function so that the complementary information between m
ultiple trees is well exploited. In this way, the fitting power of the forest is
 significantly enhanced. The global pruning is developed to reduce the model si
ze as well as the over-fitting risk. The refined model has better performance an

d smaller storage cost, as verified in extensive experiments.
****

Adaptive Region Pooling for Object Detection
Yi-Hsuan Tsai, Onur C. Hamsici, Ming-Hsuan Yang; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 731-739
Learning models for object detection is a challenging problem due to the large i ntra-class variability of objects in appearance, viewpoints, and rigidity. We ad dress this variability by a novel feature pooling method that is adaptive to seg mented regions. The proposed detection algorithm automatically discovers a diver se set of exemplars and their distinctive parts which are used to encode the reg ion structure by the proposed feature pooling method. Based on each exemplar and its parts, a regression model is learned with samples selected by a coarse regi on matching scheme. The proposed algorithm performs favorably on the PASCAL VOC 2007 dataset against existing algorithms. We demonstrate the benefits of our fea ture pooling method when compared to conventional spatial pyramid pooling featur es. We also show that object information can be transferred through exemplars fo r detected objects.
****

Discriminative and Consistent Similarities in Instance-Level Multiple Instance L earning
Mohammad Rastegari, Hannaneh Hajishirzi, Ali Farhadi; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 740-748
In this paper we present a bottom-up method to instance-level Multiple Instance Learning (MIL) that learns to discover positive instances with globally constrai ned reasoning about local pairwise similarities. We discover positive instances by optimizing for a ranking such that positive (top rank) instances are {\it hig hly and consistently similar} to each other and dissimilar to negative instances . Our approach takes advantage of a discriminative notion of pairwise similarity coupled with a structural cue in the form of a consistency metric that measures the quality of each similarity. We learn a similarity function for every pair o f instances in positive bags by how similarly they differ from instances in nega tive bags, the only certain labels in MIL. Our experiments demonstrate that our method consistently outperforms state-of-the-art MIL methods both at bag-level and instance-level predictions in standard benchmarks, image category recogniti on, and text categorization datasets.
****

MUlti-Store Tracker (MUSTer): A Cognitive Psychology Inspired Approach to Object Tracking
Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, Dacheng Tao; Proc eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2015, pp. 749-758
Variations in the appearance of a tracked object, such as changes in geometry/ph otometry, camera viewpoint, illumination, or partial occlusion, pose a major cha llenge to object tracking. Here, we adopt cognitive psychology principles to des ign a flexible representation that can adapt to changes in object appearance dur ing tracking. Inspired by the well-known Atkinson-Shiffrin Memory Model, we prop ose MUlti-Store Tracker (MUSTer), a dual-component approach consisting of short- and long-term memory stores to process target appearance memories. A powerful a nd efficient Integrated Correlation Filter (ICF) is employed in the short-term s tore for short-term tracking. The integrated long-term component, which is based on keypoint matching-tracking and RANSAC estimation, can interact with the long -term memory and provide additional information for output control. MUSTer was e xtensively evaluated on the CVPR2013 Online Object Tracking Benchmark (OOTB) and ALOV++ datasets. The experimental results demonstrated the superior performance of MUSTer in comparison with other state-of-art trackers.
****

Finding Action Tubes
Georgia Gkioxari, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 759-768
We address the problem of action detection in videos. Driven by the latest progr

ess in object detection from 2D images, we build action models using rich featur e hierarchies derived from shape and kinematic cues. We incorporate appearance a nd motion in two ways. First, starting from image region proposals we select tho se that are motion salient and thus are more likely to contain the action. This leads to a significant reduction in the number of regions being processed and al lows for faster computations. Second, we extract spatio-temporal feature represe ntations to build strong classifiers using Convolutional Neural Networks. We lin k our predictions to produce detections consistent in time, which we call action tubes. We show that our approach outperforms other techniques in the task of ac tion detection.
****

Learning a Convolutional Neural Network for Non-Uniform Motion Blur Removal
Jian Sun, Wenfei Cao, Zongben Xu, Jean Ponce; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 769-777
In this paper, we address the problem of estimating and removing non-uniform mot ion blur from a single blurry image. We propose a deep learning approach to pred icting the probabilistic distribution of motion blur at the patch level using a convolutional neural network (CNN). We further extend the candidate set of motio n kernels predicted by the CNN using carefully designed image rotations. A Marko v random field model is then used to infer a dense non-uniform motion blur field enforcing motion smoothness. Finally, motion blur is removed by a non-uniform d eblurring model using patch-level image prior. Experimental evaluations show tha t our approach can effectively estimate and remove complex non-uniform motion bl ur that is not handled well by previous approaches.
****

Complexity-Adaptive Distance Metric for Object Proposals Generation
Yao Xiao, Cewu Lu, Efstratios Tsougenis, Yongyi Lu, Chi-Keung Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 778-786
Distance metric plays a key role in grouping superpixels to produce object propo sals for object detection. We observe that existing distance metrics work primar ily for low complexity cases. In this paper, we develop a novel distance metric for grouping two superpixels in high-complexity scenarios. Combining them, a com plexity-adaptive distance measure is produced that achieves improved grouping in different levels of complexity. Our extensive experimentation shows that our me thod can achieve good results in the PASCAL VOC 2012 dataset surpassing the late st state-of-the-art methods.
****

High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild
Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 787-796
Pose and expression normalization is a crucial step to recover the canonical vie w of faces under arbitrary conditions, so as to improve the face recognition per formance. An ideal normalization method is desired to be automatic, database ind ependent and high-fidelity, where the face appearance should be preserved with l ittle artifact and information loss. However, most normalization methods fail to satisfy one or more of the goals. In this paper, we propose a High-fidelity Pos e and Expression Normalization (HPEN) method with 3D Morphable Model (3DMM) whic h can automatically generate a natural face image in frontal pose and neutral ex pression. Specifically, we firstly make a landmark marching assumption to descri be the non-correspondence between 2D and 3D landmarks caused by pose variations and propose a pose adaptive 3DMM fitting algorithm. Secondly, we mesh the whole image into a 3D object and eliminate the pose and expression variations using an identity preserving 3D transformation. Finally, we propose an inpainting method based on Possion Editing to fill the invisible region caused by self occlusion. Extensive experiments on Multi-PIE and LFW demonstrate that the proposed method significantly improves face recognition performance and outperforms state-of-th e-art methods in both constrained and unconstrained environments.
****

Transformation of Markov Random Fields for Marginal Distribution Estimation

Masaki Saito, Takayuki Okatani; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2015, pp. 797-805
This paper presents a generic method for transforming MRFs for the marginal infe
rence problem. Its major application is to downsize MRFs to speed up the computa
tion. Unlike the MAP inference, there are only classical algorithms for the marg
inal inference problem such as BP etc. that require large computational cost. Al
though downsizing MRFs should directly reduce the computational cost, there is n
o systematic way of doing this, since it is unclear how to obtain the MRF energy
 for the downsized MRFs and also how to translate the estimates of their margina
l distributions to those of the original MRFs. The proposed method resolves thes
e issues by a novel probabilistic formulation of MRF transformation. The key ide
a is to represent the joint distribution of an MRF with that of the transformed
one, in which the variables of the latter are treated as latent variables. We al
so show that the proposed method can be applied to discretization of variable sp
ace of continuous MRFs and can be used with Markov chain Monte Carlo methods. Th
e experimental results demonstrate the effectiveness of the proposed method.
****

Sparse Convolutional Neural Networks
Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, Marianna Pensky; Proceed
ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2
015, pp. 806-814
Deep neural networks have achieved remarkable performance in both image classifi
cation and object detection problems, at the cost of a large number of parameter
s and computational complexity. In this work, we show how to reduce the redundan
cy in these parameters using a sparse decomposition. Maximum sparsity is obtaine
d by exploiting both inter-channel and intra-channel redundancy, with a fine-tun
ing step that minimize the recognition loss caused by maximizing sparsity. This
procedure zeros out more than 90\% of parameters, with a drop of accuracy that i
s less than 1\% on the ILSVRC2012 dataset. We also propose an efficient sparse
matrix multiplication algorithm on CPU for Sparse Convolutional Neural Networks
(SCNN) models. Our CPU implementation demonstrates much higher efficiency than t
he off-the-shelf sparse matrix libraries, with a significant speedup realized ov
er the original dense network. In addition, we apply the SCNN model to the objec
t detection problem, in conjunction with a cascade model and sparse fully connec
ted layers, to achieve significant speedups.
****

FaceNet: A Unified Embedding for Face Recognition and Clustering
Florian Schroff, Dmitry Kalenichenko, James Philbin; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823
Despite significant recent advances in the field of face recognition [DeepFace,
DeepId2], implementing face verification and recognition efficiently at scale pr
esents serious challenges to current approaches. In this paper we present a syst
em, called FaceNet, that directly learns a mapping from face images to a compact
 Euclidean space where distances directly correspond to a measure of face simila
rity. Once this space has been produced, tasks such as face recognition, verific
ation and clustering can be easily implemented using standard techniques with Fa
ceNet embeddings as feature vectors. Our method uses a deep convolutional netwo
rk trained to directly optimize the embedding itself, rather than an intermediat
e bottleneck layer as in previous approaches. To train, we use triplets of rough
ly aligned matching / non-matching face patches generated using a novel online t
riplet mining method. The benefit of our approach is much greater representation
al efficiency: we achieve state-of-the-art face recognition performance using on
ly 128 bytes per face. On the widely used Labeled Faces in the Wild (LFW) datas
et, our system achieves a new record accuracy of 99.63%. On YouTube Faces DB it
achieves 95.12%. Our system cuts the error rate in comparison to the best publis
hed result [DeepId2+] by 30% on both datasets.
****

Cascaded Hand Pose Regression
Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, Jian Sun; Proceedings of the IE
EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 824-8

We extends the previous 2D cascaded object pose regression work [9] in two aspec ts so that it works better for 3D articulated objects. Our first contribution is 3D pose-indexed features that generalize the previous 2D parameterized features and achieve better invariance to 3D transformations. Our second contribution is a principled hierarchical regression that is adapted to the articulated object structure. It is therefore more accurate and faster. Comprehensive experiments v erify the state-of-the-art accuracy and efficiency of the proposed approach on t he challenging 3D hand pose estimation problem, on a public dataset and our new dataset.
****

Cross-Scene Crowd Counting via Deep Convolutional Neural Networks
Cong Zhang, Hongsheng Li, Xiaogang Wang, Xiaokang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 833-841
Cross-scene crowd counting is a challenging task where no laborious data annotat ion is required for counting people in new target surveillance crowd scenes unse en in the training set. The performance of most existing crowd counting methods drops significantly when they are applied to an unseen scene. To address this pr oblem, we propose a deep convolutional neural network (CNN) for crowd counting, and it is trained alternatively with two related learning objectives, crowd dens ity and crowd count. This proposed switchable learning approach is able to obtai n better local optimum for both objectives. To handle an unseen target crowd sce ne, we present a data-driven method to fine-tune the trained CNN model for the t arget scene. A new dataset including 108 crowd scenes with nearly 200,000 head a nnotations is introduced to better evaluate the accuracy of cross-scene crowd co unting methods. Extensive experiments on the proposed and another two existing d atasets demonstrate the effectiveness and reliability of our approach.
****

The Application of Two-Level Attention Models in Deep Convolutional Neural Netwo rk for Fine-Grained Image Classification
Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, Zheng Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C VPR), 2015, pp. 842-850
Fine-grained classification is challenging because categories can only be discri minated by subtle and local differences. Variances in the pose, scale or rotatio n usually make the problem more difficult. Most fine-grained classification syst ems follow the pipeline of finding foreground object or object parts (where) to extract discriminative features (what). In this paper, we propose to apply visu al attention to fine-grained classification task using deep neural network. Our pipeline integrates three types of attention: the bottom-up attention that propo se candidate patches, the object-level top-down attention that selects relevant patches to a certain object, and the part-level top-down attention that localize s discriminative parts. We combine these attentions to train domain-specific dee p nets, then use it to improve both the what and where aspects. Importantly, we avoid using expensive annotations like bounding box or part information from end -to-end. The weak supervision constraint makes our work easier to generalize. W e have verified the effectiveness of the method on the subsets of ILSVRC2012 dat aset and CUB200_2011 dataset. Our pipeline delivered significant improvements an d achieved the best accuracy under the weakest supervision condition. The perfor mance is competitive against other methods that rely on additional annotations.
****

End-to-End Integration of a Convolution Network, Deformable Parts Model and Non-Maximum Suppression
Li Wan, David Eigen, Rob Fergus; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 851-859
Deformable Parts Models and Convolutional Networks each have achieved notable pe rformance in object detection. Yet these two approaches find their strengths in complementary areas: DPMs are well-versed in object composition, modeling fine -grained spatial relationships between parts; likewise, ConvNets are adept at pr oducing powerful image features, having been discriminatively trained directly o

n the pixels. In this paper, we propose a new model that combines these two appr oaches, obtaining the advantages of each. We train this model using a new struct ured loss function that considers all bounding boxes within an image, rather tha n isolated object instances. This enables the non-maximal suppression (NMS) oper ation, previously treated as a separate post-processing stage, to be integrated into the model. This allows for discriminative training of our combined Convnet + DPM + NMS model in end-to-end fashion. We evaluate our system on PASCAL VOC 20 07 and 2011 datasets, achieving competitive results on both bench- marks.
****

A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions
Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, Andrew C. Gallagher; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp . 860-868
This paper explores two new aspects of photos and human emotions. First, we show through psychovisual studies that different people have different emotional rea ctions to the same image, which is a strong and novel departure from previous wo rk that only records and predicts a single dominant emotion for each image. Our studies also show that the same person may have multiple emotional reactions to one image. Predicting emotions in ``distributions'' instead of a single dominant emotion is important for many applications. Second, we show not only that we ca n often change the evoked emotion of an image by adjusting color tone and textur e related features but also that we can choose in which ``emotional direction" t his change occurs by selecting a target image. In addition, we present a new dat abase, Emotion6, containing distributions of emotions.
****

Neuroaesthetics in Fashion: Modeling the Perception of Fashionability
Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, Raquel Urtasun; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 15, pp. 869-877
In this paper, we analyze the fashion of clothing of a large social website. Our goal is to learn and predict how fashionable a person looks on a photograph and suggest subtle improvements the user could make to improve her/his appeal. We p ropose a Conditional Random Field model that jointly reasons about several fashi onability factors such as the type of outfit and garments the user is wearing, t he type of the user, the photograph's setting (e.g., the scenery behind the user), and the fashionability score. Importantly, our model is able to give rich fee dback back to the user, conveying which garments or even scenery she/he should c hange in order to improve fashionability. We demonstrate that our joint approach significantly outperforms a variety of intelligent baselines. We additionally c ollected a novel heterogeneous dataset with 144,169 user posts containing divers e image, textual and meta information which can be exploited for our task. We al so provide a detailed analysis of the data, showing different outfit trends and fashionability scores across the globe and across a span of 6 years.
****

Part-Based Modelling of Compound Scenes From Images
Anton van den Hengel, Chris Russell, Anthony Dick, John Bastian, Daniel Pooley, Lachlan Fleming, Lourdes Agapito; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 878-886
We propose a method to recover the structure of a compound scene from multiple s ilhouettes. Structure is expressed as a collection of 3D primitives chosen from a pre-defined library, each with an associated pose. This has several advantages over a volume or mesh representation both for estimation and the utility of the recovered model. The main challenge in recovering such a model is the combinato rial number of possible arrangements of parts. We address this issue by exploiti ng the intrinsic structure and sparsity of the problem, and show that our method scales to scenes constructed from large libraries of parts.
****

Efficient Parallel Optimization for Potts Energy With Hierarchical Fusion
Olga Veksler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 887-895

Potts energy frequently occurs in computer vision applications. We present an efficient parallel method for optimizing Potts energy based on the extension of hierarchical fusion algorithm. Unlike previous parallel graph-cut based optimization algorithms, our approach has optimality bounds even after a single iteration over all labels, i.e. after solving only k-1 max-flow problems, where k is the number of labels. This is perhaps the minimum number of max-flow problems one has to solve to obtain a solution with optimality guarantees. Our approximation factor is O(log k). Although this is not as good as the factor of 2 approximation of the well known expansion algorithm, we achieve very good results in practice. In particular, we found that the results of our algorithm after one iteration are always better than the results after one iteration of the expansion algorithm. We demonstrate experimentally the computational advantages of our parallel implementation on the problem of stereo correspondence, achieving a factor of 1.5 to 2.6 speedup compared to the serial implementation. These results were obtained with a small number of processors. The expected speedups with a larger number of processors are greater.
****

Pooled Motion Features for First-Person Videos
Michael S. Ryoo, Brandon Rothrock, Larry Matthies; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 896-904
In this paper, we present a new feature representation for first-person videos. In first-person video understanding (e.g., activity recognition), it is very important to capture both entire scene dynamics (i.e., egomotion) and salient local motion observed in videos. We describe a representation framework based on time series pooling, which is designed to abstract short-term/long-term changes in feature descriptor elements. The idea is to keep track of how descriptor values are changing over time and summarize them to represent motion in the activity video. The framework is general, handling any types of per-frame feature descriptors including conventional motion descriptors like histogram of optical flows (HOF) as well as appearance descriptors from more recent convolutional neural networks (CNN). We experimentally confirm that our approach clearly outperforms previous feature representations including bag-of-visual-words and improved Fisher vector (IFV) when using identical underlying feature descriptors. We also confirm that our feature representation has superior performance to existing state-of-the-art features like local spatio-temporal features and Improved Trajectory Features (originally developed for 3rd-person videos) when handling first-person videos. Multiple first-person activity datasets were tested under various settings to confirm these findings.
****

Functional Correspondence by Matrix Completion
Artiom Kovnatsky, Michael M. Bronstein, Xavier Bresson, Pierre Vandergheynst; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 905-914
In this paper, we consider the problem of finding dense intrinsic correspondence between manifolds using the recently introduced functional framework. We pose the functional correspondence problem as matrix completion with manifold geometric structure and inducing functional localization with the L1 norm. We discuss efficient numerical procedures for the solution of our problem. Our method compares favorably to the accuracy of state-of-the-art correspondence algorithms on non-rigid shape matching benchmarks, and is especially advantageous in settings when only scarce data is available.
****

Elastic-Net Regularization of Singular Values for Robust Subspace Learning
Eunwoo Kim, Minsik Lee, Songhwai Oh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 915-923
Learning a low-dimensional structure plays an important role in computer vision. Recently, a new family of methods, such as l1 minimization and robust principal component analysis, has been proposed for low-rank matrix approximation problems and shown to be robust against outliers and missing data. But these methods often require heavy computational load and can fail to find a solution when highly

corrupted data are presented. In this paper, an elastic-net regularization base
d low-rank matrix factorization method for subspace learning is proposed. The pr
oposed method finds a robust solution efficiently by enforcing a strong convex c
onstraint to improve the algorithm's stability while maintaining the low-rank pr
operty of the solution. It is shown that any stationary point of the proposed al
gorithm satisfies the Karush-Kuhn-Tucker optimality conditions. The proposed met
hod is applied to a number of low-rank matrix approximation problems to demonstr
ate its efficiency in the presence of heavy corruptions and to show its effectiv
eness and robustness compared to the existing methods.
****

Hardware Compliant Approximate Image Codes
Da Kuang, Alex Gittens, Raffay Hamid; Proceedings of the IEEE Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2015, pp. 924-932
In recent years, several feature encoding schemes for the bags-of-visual-words m
odel have been proposed. While most of these schemes produce impressive results,
 they all share an important limitation: their high computational complexity mak
es it challenging to use them for large-scale problems. In this work, we propose
 an approximate locality-constrained encoding scheme that offers significantly b
etter computational efficiency (~40x) than its exact counterpart, with comparabl
e classification accuracy. Using the perturbation analysis of least-squares prob
lems, we present a formal approximation error analysis of our approach, which he
lps distill the intuition behind the robustness of our method. We present a thor
ough set of empirical analyses on multiple standard data-sets, to assess the cap
ability of our encoding scheme for its representational as well as discriminativ
e accuracy.
****

Photometric Refinement of Depth Maps for Multi-Albedo Objects
Avishek Chatterjee, Venu Madhav Govindu; Proceedings of the IEEE Conference on C
omputer Vision and Pattern Recognition (CVPR), 2015, pp. 933-941
In this paper, we propose a novel uncalibrated photometric method for refining d
epth maps of multi-albedo objects obtained from consumer depth cameras like Kine
ct. Existing uncalibrated photometric methods either assume that the object has
constant albedo or rely on segmenting images into constant albedo regions. The m
ethod of this paper does not require the constant albedo assumption and we belie
ve it is the first work of its kind to handle objects with arbitrary albedo unde
r uncalibrated illumination. We first robustly estimate a rank 3 approximation o
f the observed brightness matrix using an iterative reweighting method. Subseque
ntly, we factorize this rank reduced brightness matrix into the corresponding li
ghting, albedo and surface normal components. The proposed factorization is show
n to be convergent. We experimentally demonstrate the value of our approach by p
resenting highly accurate three-dimensional reconstructions of a wide variety of
 objects. Additionally, since any photometric method requires a radiometric cali
bration of the camera used, we also present a direct radiometric calibration tec
hnique for the infra-red camera of the structured-light stereo depth scanner. Un
like existing methods, this calibration technique does not depend on a known cal
ibration object or on the properties of the scene illumination used.
****

Predicting the Future Behavior of a Time-Varying Probability Distribution
Christoph H. Lampert; Proceedings of the IEEE Conference on Computer Vision and
Pattern Recognition (CVPR), 2015, pp. 942-950
We study the problem of predicting the future, though only in the probabilistic
sense of estimating a future state of a time-varying probability distribution. T
his is not only an interesting academic problem, but solving this extrapolation
problem also has many practical application, e.g. for training classifiers that
have to operate under time-varying conditions. Our main contribution is a method
 for predicting the next step of the time-varying distribution from a given sequ
ence of sample sets from earlier time steps. For this we rely on two recent mach
ine learning techniques: embedding probability distributions into a reproducing
kernel Hilbert space, and learning operators by vector-valued regression. We ill
ustrate the working principles and the practical usefulness of our method by exp

eriments on synthetic and real data. We also highlight an exemplary application: training a classifier in a domain adaptation setting without having access to examples from the test time distribution at training time.

****

Classifier Based Graph Construction for Video Segmentation

Anna Khoreva, Fabio Galasso, Matthias Hein, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 951-960

Video segmentation has become an important and active research area with a large diversity of proposed approaches. Graph-based methods, enabling topperformance on recent benchmarks, consist of three essential components: 1. powerful features account for object appearance and motion similarities; 2. spatio-temporal neighborhoods of pixels or superpixels (the graph edges) are modeled using a combination of those features; 3. video segmentation is formulated as a graph partitioning problem. While a wide variety of features have been explored and various graph partition algorithms have been proposed, there is surprisingly little research on how to construct a graph to obtain the best video segmentation performance. This is the focus of our paper. We propose to combine features by means of a classifier, use calibrated classifier outputs as edge weights and define the graph topology by edge selection. By learning the graph (without changes to the graph partitioning method), we improve the results of the best performing video segmentation algorithm by 6% on the challenging VSB100 benchmark, while reducing its runtime by 55%, as the learnt graph is much sparser.

****

ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, Juan Carlos Niebles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 961-970

In spite of many dataset efforts for human action recognition, current computer vision algorithms are still severely limited in terms of the variability and complexity of the actions that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on simple actions and movements occurring on manually trimmed videos. In this paper, we introduce ActivityNet: a new large-scale video benchmark for human activity understanding. Our new benchmark aims at covering a wide range of complex human activities that are of interest to people in their daily living. In its current version, ActivityNet provides samples from 203 activity categories with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 hours of video. We illustrate three scenarios in which ActivityNet can be used to benchmark and compare algorithms for human activity understanding: untrimmed video classification, trimmed activity classification and activity detection.

****

Mid-Level Deep Pattern Mining

Yao Li, Lingqiao Liu, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 971-980

Mid-level visual element discovery aims to find clusters of image patches that are both representative and discriminative. In this work, we study this problem from the prospective of pattern mining while relying on the recently popularized Convolutional Neural Networks (CNNs). Specifically, we find that for an image patch, activation extracted from the first fully-connected layer of a CNN have two appealing properties which enable its seamless integration with pattern mining. Patterns are then discovered from a large number of CNN activations of image patches through the well-known association rule mining. When we retrieve and visualize image patches with the same pattern (See Fig. 1), surprisingly, they are not only visually similar but also semantically consistent. We apply our approach to scene and object classification tasks, and demonstrate that our approach outperforms all previous works on mid-level visual element discovery by a sizeable margin with far fewer elements being used. Our approach also outperforms or matches recent works using CNN for these tasks. Source code of the complete system is

available online.
****
Prediction of Search Targets From Fixations in Open-World Settings
Hosnieh Sattar, Sabine Muller, Mario Fritz, Andreas Bulling; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 981-990
Previous work on predicting the target of visual search from human fixations only considered closed-world settings in which training labels are available and predictions are performed for a known set of potential targets. In this work we go beyond the state of the art by studying search target prediction in an open-world setting in which we no longer assume that we have fixation data to train for the search targets. We present a dataset containing fixation data of 18 users searching for natural images from three image categories within synthesised image collages of about 80 images. In a closed-world baseline experiment we show that we can predict the correct target image out of a candidate set of five images. We then present a new problem formulation for search target prediction in the open-world setting that is based on learning compatibilities between fixations and potential targets.
****
Understanding Image Representations by Measuring Their Equivariance and Equivalence
Karel Lenc, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 991-999
Despite the importance of image representations such as histograms of oriented gradients and deep Convolutional Neural Networks (CNN), our theoretical understanding of them remains limited. Aiming at filling this gap, we investigate three key mathematical properties of representations: equivariance, invariance, and equivalence. Equivariance studies how transformations of the input image are encoded by the representation, invariance being a special case where a transformation has no effect. Equivalence studies whether two representations, for example two different parametrisations of a CNN, capture the same visual information or not. A number of methods to establish these properties empirically are proposed, including introducing transformation and stitching layers in CNNs. These methods are then applied to popular representations to reveal insightful aspects of their structure, including clarifying at which layers in a CNN certain geometric invariances are achieved. While the focus of the paper is theoretical, direct applications to structured-output regression are demonstrated too.
****
Effective Learning-Based Illuminant Estimation Using Simple Features
Dongliang Cheng, Brian Price, Scott Cohen, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1000-1008
Illumination estimation is the process of determining the chromaticity of the illumination in an imaged scene in order to remove undesirable color casts through white-balancing. While computational color constancy is a well-studied topic in computer vision, it remains challenging due to the ill-posed nature of the problem. One class of techniques relies on low-level statistical information in the image color distribution and works under various assumptions (e.g. Grey-World, White-Patch, etc). These methods have an advantage that they are simple and fast, but often do not perform well. More recent state-of-the-art methods employ learning-based techniques that produce better results, but often rely on complex features and have long evaluation and training times. In this paper, we present a learning-based method based on four simple color features and show how to use this with an ensemble of regression trees to estimate the illumination. We demonstrate that our approach is not only faster than existing learning-based methods in terms of both evaluation and training time, but also gives the best results reported to date on modern color constancy data sets.
****
PAIGE: PAirwise Image Geometry Encoding for Improved Efficiency in Structure-From-Motion

Johannes L. Schonberger, Alexander C. Berg, Jan-Michael Frahm; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1 009-1018
Large-scale Structure-from-Motion systems typically spend major computational ef fort on pairwise image matching and geometric verification in order to discover connected components in large-scale, unordered image collections. In recent year s, the research community has spent significant effort on improving the efficien cy of this stage. In this paper, we present a comprehensive overview of various state-of-the-art methods, evaluating and analyzing their performance. Based on t he insights of this evaluation, we propose a learning-based approach, the PAirwi se Image Geometry Encoding (PAIGE), to efficiently identify image pairs with sce ne overlap without the need to perform exhaustive putative matching and geometri c verification. PAIGE achieves state-of-the-art performance and integrates well into existing Structure-from-Motion pipelines.
****

Dense, Accurate Optical Flow Estimation With Piecewise Parametric Model
Jiaolong Yang, Hongdong Li; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2015, pp. 1019-1027
This paper proposes a simple method for estimating dense and accurate optical fl ow field. It revitalizes an early idea of piecewise parametric flow model. A key innovation is that, we fit a flow field piecewise to a variety of parametric mo dels, where the domain of each piece (i.e., each piece's shape, position and siz e) is determined adaptively, while at the same time maintaining a global inter-p iece flow continuity constraint. We achieve this by a multi-model fitting scheme via energy minimization. Our energy takes into account both the piecewise const ant model assumption and the flow field continuity constraint, enabling the prop osed method to effectively handle both homogeneous motions and complex motions. The experiments on three public optical flow benchmarks (KITTI, MPI Sintel, and Middlebury) show the superiority of our method compared with the state of the ar t: it achieves top-tier performances on all the three benchmarks.
****

Single-Image Estimation of the Camera Response Function in Near-Lighting
Pedro Rodrigues, Joao P. Barreto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1028-1036
The camera response function (CRF) relates quantised image pixel values with phy sical incoming light. This paper describes a method to estimate the CRF from a s ingle image of a general two-coloured surface for which the albedo ratio between the coloured regions is known a priori. While other radiometric calibration met hods either use multiple frames or require the light to be infinitely distant, t he algorithm herein proposed makes no assumptions about lighting conditions and can handle cameras with strong vignetting. Although the approach is generic, in the sense that can be applied to any camera system, the method is particularly w ell suited for determining the CRF of near-lighting endoscopes in the operating room. This is a very pertinent problem for which no practical, effective solutio ns have been proposed. The robustness, repeatability, and accuracy of the algori thm is experimentally validated in real images acquired with different endoscopi c set-ups.
****

Multispectral Pedestrian Detection: Benchmark Dataset and Baseline
Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp . 1037-1045
With the increasing interest in pedestrian detection, pedestrian datasets have a lso been the subject of research in the past decades. However, most existing dat asets focus on a color channel, while a thermal channel is helpful for detection even in a dark environment. With this in mind, we propose a multispectral pedes trian dataset which provides well aligned color-thermal image pairs, captured by beam splitter-based special hardware. The color-thermal dataset is as large as previous color-based datasets and provides dense annotations including temporal correspondences. With this dataset, we introduce multispectral ACF, which is an

extension of aggregated channel features (ACF) to simultaneously handle color-th
ermal image pairs. Multispectral ACF reduces the average miss rate of ACF by 15%
, and achieves another breakthrough in the pedestrian detection task.
****

A Low-Dimensional Step Pattern Analysis Algorithm With Application to Multimodal
 Retinal Image Registration
Jimmy Addison Lee, Jun Cheng, Beng Hai Lee, Ee Ping Ong, Guozhen Xu, Damon Wing
Kee Wong, Jiang Liu, Augustinus Laude, Tock Han Lim; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1046-1053
Existing feature descriptor-based methods on retinal image registration are main
ly based on scale-invariant feature transform (SIFT) or partial intensity invari
ant feature descriptor (PIIFD). While these descriptors are often being exploite
d, they do not work very well upon unhealthy multimodal images with severe disea
ses. Additionally, the descriptors demand high dimensionality to adequately repr
esent the features of interest. The higher the dimensionality, the greater the c
onsumption of resources (e.g. memory space). To this end, this paper introduces
a novel registration algorithm coined low-dimensional step pattern analysis (LoS
PA), tailored to achieve low dimensionality while providing sufficient distincti
veness to effectively align unhealthy multimodal image pairs. The algorithm loca
tes hypotheses of robust corner features based on connecting edges from the edge
 maps, mainly formed by vascular junctions. This method is insensitive to intens
ity changes, and produces uniformly distributed features and high repeatability
across the image domain. The algorithm continues with describing the corner feat
ures in a rotation invariant manner using step patterns. These customized step p
atterns are robust to non-linear intensity changes, which are well-suited for mu
ltimodal retinal image registration. Apart from its low dimensionality, the LoSP
A algorithm achieves about two-fold higher success rate in multimodal registrati
on on the dataset of severe retinal diseases when compared to the top score amon
g state-of-the-art algorithms.
****

Bilinear Heterogeneous Information Machine for RGB-D Action Recognition
Yu Kong, Yun Fu; Proceedings of the IEEE Conference on Computer Vision and Patte
rn Recognition (CVPR), 2015, pp. 1054-1062
This paper proposes a novel approach to action recognition from RGB-D cameras, i
n which depth features and RGB visual features are jointly used. Rich heterogene
ous RGB and depth data are effectively compressed and projected to a learned sha
red space, in order to reduce noise and capture useful information for recogniti
on. Knowledge from various sources can then be shared with others in the learned
 space to learn cross-modal features. This guides the discovery of valuable info
rmation for recognition. To capture complex spatiotemporal structural relationsh
ips in visual and depth features, we represent both RGB and depth data in a matr
ix form. We formulate the recognition task as a low-rank bilinear model composed
 of row and column parameter matrices. The rank of the model parameter is minimi
zed to build a low-rank classifier, which is beneficial for improving the genera
lization power. The proposed method is extensively evaluated on two public RGB-D
 action datasets, and achieves state-of-the-art results. It also shows promising
 results if RGB or depth data are missing in training or testing procedure.
****

MRF Optimization by Graph Approximation
Wonsik Kim, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2015, pp. 1063-1071
Graph cuts-based algorithms have achieved great success in energy minimization f
or many computer vision applications. These algorithms provide approximated solu
tions for multi-label energy functions via move-making approach. This approach f
uses the current solution with a proposal to generate a lower-energy solution. T
hus, generating the appropriate proposals is necessary for the success of the mo
ve-making approach. However, not much research efforts has been done on the gene
ration of ``good'' proposals, especially for non-metric energy functions. In thi
s paper, we propose an application-independent and energy-based approach to gene
rate ``good'' proposals. With these proposals, we present a graph cuts-based mov

e-making algorithm called GA-fusion (fusion with graph approximation-based propo sals). Extensive experiments support that our proposal generation is effective a cross different classes of energy functions. The proposed algorithm outperforms others both on real and synthetic problems.
****

SALICON: Saliency in Context
Ming Jiang, Shengsheng Huang, Juanyong Duan, Qi Zhao; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1072-1080
Saliency in Context (SALICON) is an ongoing effort that aims at understanding an d predicting visual attention. This paper presents a new method to collect large -scale human data during natural explorations on images. While current datasets present a rich set of images and task-specific annotations such as category labe ls and object segments, this work focuses on recording and logging how humans sh ift their attention during visual exploration. The goal is to offer new possibil ities to (1) complement task-specific annotations to advance the ultimate goal i n visual understanding, and (2) understand visual attention and learn saliency m odels, all with human attentional data at a much larger scale. We designed a mo use-contingent multi-resolutional paradigm based on neurophysiological and psych ophysical studies of peripheral vision, to simulate the natural viewing behavior of humans. The new paradigm allowed using a general-purpose mouse instead of an eye tracker to record viewing behaviors, thus enabling large-scale data collect ion. The paradigm was validated with controlled laboratory as well as large-scal e online data. We report in this paper a proof-of-concept SALICON dataset of hum an "free-viewing" data on 10,000 images from the Microsoft COCO (MS COCO) datase t with rich contextual information. We evaluated the use of the collected data i n the context of saliency prediction, and demonstrated them a good source as gro und truth for the evaluation of saliency algorithms.
****

Weakly Supervised Object Detection With Convex Clustering
Hakan Bilen, Marco Pedersoli, Tinne Tuytelaars; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1081-1089
Weakly supervised object detection, is a challenging task, where the training pr ocedure involves learning at the same time both, the model appearance and the ob ject location in each image. The classical approach to solve this problem is to consider the location of the object of interest in each image as a latent variab le and minimize the loss generated by such latent variable during learning. How ever, as learning appearance and localization are two interconnected tasks, the optimization is not convex and the procedure can easily get stuck in a poor loca l minimum, the algorithm "misses" the object in some images. In this paper, we h elp the optimization to get close to the global minimum by enforcing a "soft" si milarity between each possible location in the image and a reduced set of "exemp lars", or clusters, learned with a convex formulation in the training images. T he help is effective because it comes from a different and smooth source of info rmation that is not directly connected with the main task. Results show that ou r method improves a strong baseline based on convolutional neural network featur es by more than 4 points without any additional features or extra computation at testing time but only adding a small increment of the training time due to the convex clustering.
****

Interleaved Text/Image Deep Mining on a Very Large-Scale Radiology Database
Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, Ronald M. Summers; Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1090-1099
Despite tremendous progress in computer vision, effective learning on very large -scale (>100K patients) medical image databases has been vastly hindered. We pre sent an interleaved text/image deep learning system to extract and mine the sema ntic interactions of radiology images and reports from a national research hospi tal's picture archiving and communication system. Instead of using full 3D medic al volumes, we focus on a collection of representative ~216K 2D key images/slice s (selected by clinicians for diagnostic reference) with text-driven scalar and

vector labels. Our system interleaves between unsupervised learning (e.g., laten t Dirichlet allocation, recurrent neural net language models) on document- and s entence-level texts to generate semantic labels and supervised learning via deep convolutional neural networks (CNNs) to map from images to label spaces. Diseas e-related key words can be predicted for radiology images in a retrieval manner. We have demonstrated promising quantitative and qualitative results. The large-scale datasets of extracted key images and their categorization, embedded vector labels and sentence descriptions can be harnessed to alleviate the deep learnin g "data-hungry" obstacle in the medical domain.
****

Learning Semantic Relationships for Better Action Retrieval in Images
Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang So ng, Samy Bengio, Charles Rosenberg, Li Fei-Fei; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1100-1109
Human actions capture a wide variety of interactions between people and objects. As a result, the set of possible actions is extremely large and it is difficult to obtain sufficient training examples for all actions. However, we could compe nsate for this sparsity in supervision by leveraging the rich semantic relations hip between different actions. A single action is often composed of other smalle r actions and is exclusive of certain others. We need a method which can reason about such relationships and extrapolate unobserved actions from known actions. Hence, we propose a novel neural network framework which jointly extracts the re lationship between actions and uses them for training better action retrieval mo dels. Our model incorporates linguistic, visual and logical consistency based cu es to effectively identify these relationships. We train and test our model on a largescale image dataset of human actions. We show a significant improvement in mean AP compared to different baseline methods including the HEX-graph approach from Deng et al.
****

Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition
Yong Du, Wei Wang, Liang Wang; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2015, pp. 1110-1118
Human actions can be represented by the trajectories of skeleton joints. Traditi onal methods generally model the spatial structure and temporal dynamics of huma n skeleton with hand-crafted features and recognize human actions by well-design ed classifiers. In this paper, considering that recurrent neural network (RNN) c an model the long-term contextual information of temporal sequences well, we pro pose an end-to-end hierarchical RNN for skeleton based action recognition. Inste ad of taking the whole skeleton as the input, we divide the human skeleton into five parts according to human physical structure, and then separately feed them to five subnets. As the number of layers increases, the representations extracte d by the subnets are hierarchically fused to be the inputs of higher layers. The final representations of the skeleton sequences are fed into a single-layer per ceptron, and the temporally accumulated output of the perceptron is the final de cision. We compare with five other deep RNN architectures derived from our model to verify the effectiveness of the proposed network, and also compare with seve ral other methods on three publicly available datasets. Experimental results dem onstrate that our model achieves the state-of-the-art performance with high comp utational efficiency.
****

Depth and Surface Normal Estimation From Monocular Images Using Regression on De ep Features and Hierarchical CRFs
Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, Mingyi He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp . 1119-1127
Predicting the depth (or surface normal) of a scene from single monocular color images is a challenging task. This paper tackles this challenging and essentiall y under-determined problem by regression on deep convolutional neural network (D CNN) features, combined with a post-processing refining step using conditional r andom fields(CRF). Our framework works at two levels, super-pixel level and pixe

l level. First, we design a DCNN model to learn the mapping from multi-scale ima ge patches to depth or surface normal values at the super-pixel level. Second, t he estimated super-pixel depth or surface normal is refined to the pixel level b y exploiting various potentials on the depth or surface normal map, which includ es a data term, a smoothness term among super-pixels and an auto-regression term characterizing the local structure of the estimation map. The inference problem can be efficiently solved because it admits a closed-form solution. Experiments on the Make3D and NYU Depth V2 datasets show competitive results compared with recent state-of-the-art methods.

****

Discriminative Shape From Shading in Uncalibrated Illumination
Stephan R. Richter, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1128-1136
Estimating surface normals from just a single image is challenging. To simplify the problem, previous work focused on special cases, including directional light ing, known reflectance maps, etc., making shape from shading impractical outside the lab. To cope with more realistic settings, shading cues need to be combined and generalized to natural illumination. This significantly increases the compl exity of the approach, as well as the number of parameters that require tuning. Enabled by a new large-scale dataset for training and analysis, we address this with a discriminative learning approach to shape from shading, which uses regres sion forests for efficient pixel-independent prediction and fast learning. Von M ises-Fisher distributions in the leaves of each tree enable the estimation of su rface normals. To account for their expected spatial regularity, we introduce sp atial features, including texton and silhouette features. The proposed silhouett e features are computed from the occluding contours of the surface and provide s cale-invariant context. Aside from computational efficiency, they enable good ge neralization to unseen data and importantly allow for a robust estimation of the reflectance map, extending our approach to the uncalibrated setting. Experiment s show that our discriminative approach outperforms state-of-the-art methods on synthetic and real-world datasets.

****

Multi-Manifold Deep Metric Learning for Image Set Classification
Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, Jie Zhou; Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1137 -1145
In this paper, we propose a multi-manifold deep metric learning (MMDML) method f or image set classification, which aims to recognize an object of interest from a set of image instances captured from varying viewpoints or under varying illum inations. Motivated by the fact that manifold can be effectively used to model t he nonlinearity of samples in each image set and deep learning has demonstrated superb capability to model the nonlinearity of samples, we propose a MMDML metho d to learn multiple sets of nonlinear transformations, one set for each object c lass, to nonlinearly map multiple sets of image instances into a shared feature subspace, under which the manifold margin of different class is maximized, so th at both discriminative and class-specific information can be exploited, simultan eously. Our method achieves the state-of-the-art performance on five widely used datasets.

****

Target Identity-Aware Network Flow for Online Multiple Target Tracking
Afshin Dehghan, Yicong Tian, Philip H. S. Torr, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 11 46-1154
In this paper we show that multiple object tracking (MOT) can be formulated in a framework, where the detection and data-association are performed simultaneousl y. Our method allows us to overcome the confinements of data association based M OT approaches; where the performance is dependent on the object detection result s provided at input level. At the core of our method lies structured learning wh ich learns a model for each target and infers the best location of all targets s imultaneously in a video clip. The inference of our structured learning is done

through a new Target Identity-aware Network Flow (TINF), where each node in the network encodes the probability of each target identity belonging to that node. The proposed Lagrangian relaxation optimization finds the high quality solution to the network. During optimization a soft spatial constraint is enforced betwee n the nodes of the graph which helps reducing the ambiguity caused by nearby tar gets with similar appearance in crowded scenarios. We show that automatically de tecting and tracking targets in a single framework can help resolve the ambiguit ies due to frequent occlusion and heavy articulation of targets. Our experiments involve challenging yet distinct datasets and show that our method can achieve results better than the state-of-art.

****

Adaptive As-Natural-As-Possible Image Stitching

Chung-Ching Lin, Sharathchandra U. Pankanti, Karthikeyan Natesan Ramamurthy, Ale ksandr Y. Aravkin; Proceedings of the IEEE Conference on Computer Vision and Pat tern Recognition (CVPR), 2015, pp. 1155-1163

The goal of image stitching is to create natural-looking mosaics free of artifac ts that may occur due to relative camera motion, illumination changes, and optic al aberrations. In this paper, we propose a novel stitching method, that uses a smooth stitching field over the entire target image, while accounting for all th e local transformation variations. Computing the warp is fully automated and use s a combination of local homography and global similarity transformations, both of which are estimated with respect to the target. We mitigate the perspective d istortion in the non-overlapping regions by linearizing the homography and slowl y changing it to the global similarity. The proposed method is easily generalize d to multiple images, and allows one to automatically obtain the best perspectiv e in the panorama. It is also more robust to parameter selection, and hence more automated compared with state-of-the-art methods. The benefits of this method a re demonstrated using a variety of challenging cases.

****

EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow

Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, Cordelia Schmid; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 , pp. 1164-1172

We propose a novel approach for optical flow estimation, targeted at large displ acements with significant occlusions. It consists of two steps: i) dense matchin g by edge-preserving interpolation from a sparse set of matches; ii) variational energy minimization initialized with the dense matches. The sparse-to-dense int erpolation relies on an appropriate choice of the distance, namely an edge-aware geodesic distance. This distance is tailored to handle occlusions and motion bo undaries - two common and difficult issues for optical flow computation. We also propose an approximation scheme for the geodesic distance to allow fast computa tion without loss of performance. Subsequent to the dense interpolation step, st andard one-level variational energy minimization is carried out on the dense mat ches to obtain the final flow estimation. The proposed approach, called Edge-Pre serving Interpolation of Correspondences (EpicFlow) is fast and robust to large displacements. It significantly outperforms the state of the art on MPI-Sintel a nd performs on par on Kitti and Middlebury.

****

Learning Coarse-to-Fine Sparselets for Efficient Object Detection and Scene Clas sification

Gong Cheng, Junwei Han, Lei Guo, Tianming Liu; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1173-1181

Part model-based methods have been successfully applied to object detection and scene classification and have achieved state-of-the-art results. More recently t he "sparselets" work [1-3] were introduced to serve as a universal set of shared basis learned from a large number of part detectors, resulting in notable speed up. Inspired by this framework, in this paper, we propose a novel scheme to trai n more effective sparselets with a coarse-to-fine framework. Specifically, we fi rst train coarse sparselets to exploit the redundancy existing among part detect ors by using an unsupervised single-hidden layer auto-encoder. Then, we simultan

eously train fine sparselets and activation vectors using a supervised single-hi
dden-layer neural network, in which sparselets training and discriminative activ
ation vectors learning are jointly embedded into a unified framework. In order t
o adequately explore the discriminative information hidden in the part detectors
 and to achieve sparsity, we propose to optimize a new discriminative objective
function by imposing L0-norm sparsity constraint on the activation vectors. By u
sing the proposed framework, promising results for multi-class object detection
and scene classification are achieved on PASCAL VOC 2007, MIT Scene-67, and UC M
erced Land Use datasets, compared with the existing sparselets baseline methods.
****

Continuous Visibility Feature
Guilin Liu, Yotam Gingold, Jyh-Ming Lien; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1182-1190
In this work, we propose a new type of visibility measurement named Continuous V
isibility Feature (CVF). We say that a point q on the mesh is continuously visib
le from another point p if there exists a geodesic path connecting p and q that
is entirely visible by p. In order to efficiently estimate the continuous visibi
lity for all the vertices in a model, we propose two approaches that use specifi
c CVF properties to avoid exhaustive visibility tests. CVF is then measured as t
he area of the continuously visible region. With this stronger visibility measur
e, we show that CVF better encodes the surface and part information of mesh than
 the tradition line-of-sight based visibility. For example, we show that existin
g segmentation algorithms can generate better segmentation results using CVF and
 its variants than using other visibility-based shape descriptors, such as shape
 diameter function. Similar to visibility and other mesh surface features, conti
nuous visibility would have many applications.
****

FlowWeb: Joint Image Set Alignment by Weaving Consistent, Pixel-Wise Corresponde
nces
Tinghui Zhou, Yong Jae Lee, Stella X. Yu, Alyosha A. Efros; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1191
-1200
Given a set of poorly aligned images of the same visual concept without any anno
tations, we propose an algorithm to jointly bring them into pixel-wise correspon
dence by estimating a FlowWeb representation of the image set. FlowWeb is a full
y-connected correspondence flow graph with each node representing an image, and
each edge representing the correspondence flow field between a pair of images, i
.e. a vector field indicating how each pixel in one image can find a correspondi
ng pixel in the other image. Correspondence flow is related to optical flow but
allows for correspondences between visually dissimilar regions if there is evide
nce they correspond transitively on the graph. Our algorithm starts by initializ
ing all edges of this complete graph with an off-the-shelf, pairwise flow method
. We then iteratively update the graph to force it to be more self- consistent.
Once the algorithm converges, dense, globally-consistent correspondences can be
read off the graph. Our results suggest that FlowWeb improves alignment accuracy
 over previous pairwise as well as joint alignment methods.
****

Unsupervised Object Discovery and Localization in the Wild: Part-Based Matching
With Bottom-Up Region Proposals
Minsu Cho, Suha Kwak, Cordelia Schmid, Jean Ponce; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1201-1210
This paper addresses unsupervised discovery and localization of dominant objects
 from a noisy image collection with multiple object classes. The setting of this
 problem is fully unsupervised, without even image-level annotations or any assu
mption of a single dominant class. This is far more general than typical coloca
lization, cosegmentation, or weakly-supervised localization tasks. We tackle th
e discovery and localization problem using a part-based region matching approach
: We use off-the-shelf region proposals to form a set of candidate bounding boxe
s for objects and object parts. These regions are efficiently matched across ima
ges using a probabilistic Hough transform that evaluates the confidence for each

candidate correspondence considering both appearance and spatial consistency. Dominant objects are discovered and localized by comparing the scores of candidate regions and selecting those that stand out over other regions containing them. Extensive experimental evaluations on standard benchmarks demonstrate that the proposed approach significantly outperforms the current state of the art in colocalization, and achieves robust object discovery in challenging mixed-class datasets.

****

Supervised Descriptor Learning for Multi-Output Regression

Xiantong Zhen, Zhijie Wang, Mengyang Yu, Shuo Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1211-1218

Descriptor learning has recently drawn increasing attention in computer vision, Existing algorithms are mainly developed for classification rather than for regression which however has recently emerged as a powerful tool to solve a broad range of problems, e.g., head pose estimation. In this paper, we propose a novel supervised descriptor learning (SDL) algorithm to establish a discriminative and compact feature representation for multi-output regression. By formulating as generalized low-rank approximations of matrices with a supervised manifold regularization (SMR), the SDL removes irrelevant and redundant information from raw features by transforming into a low-dimensional space under the supervision of multivariate targets. The obtained discriminative while compact descriptor largely reduces the variability and ambiguity in multi-output regression, and therefore enables more accurate and efficient multivariate estimation. We demonstrate the effectiveness of the proposed SDL algorithm on a representative multi-output regression task: head pose estimation using the benchmark Pointing'04 dataset. Experimental results show that the SDL can achieve high pose estimation accuracy and significantly outperforms state-of-the-art algorithms by an error reduction up to 27.5%. The proposed SDL algorithm provides a general descriptor learning framework in a supervised way for multi-output regression which can largely boost the performance of existing multi-output regression tasks.

****

A Statistical Model of Riemannian Metric Variation for Deformable Shape Analysis

Andrea Gasparetto, Andrea Torsello; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1219-1228

The analysis of deformable 3D shape is often cast in terms of the shape's intrinsic geometry due to its invariance to a wide range of non-rigid deformations. However, object's plasticity in non-rigid transformation often results in transformations that are not completely isometric in the surface's geometry and whose mode of deviation from isometry is an identifiable characteristic of the shape and its deformation modes. In this paper, we propose a novel generative model of the variations of the intrinsic metric of deformable shapes, based on the spectral decomposition of the Laplace-Beltrami operator. To this end, we assume two independent models for the eigenvectors and the eigenvalues of the graph-Laplacian of a 3d mesh which are learned in a supervised way from a set of shapes belonging to the same class. We show how this model can be efficiently learned given a set of 3D meshes, and evaluate the performance of the resulting generative model in shape classification and retrieval tasks. Comparison with state-of-the-art solutions for these problems confirm the validity of the approach.

****

Temporally Coherent Interpretations for Long Videos Using Pattern Theory

Fillipe Souza, Sudeep Sarkar, Anuj Srivastava, Jingyong Su; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1229-1237

Graph-theoretical methods have successfully provided semantic and structural interpretations of images and videos. A recent paper introduced a pattern-theoretic approach that allows construction of flexible graphs for representing interactions of actors with objects and inference is accomplished by an efficient annealing algorithm. Actions and objects are termed generators and their interactions are termed bonds; together they form high-probability configurations, or interpretations, of observed scenes. This work and other structural methods have general

ly been limited to analyzing short videos involving isolated actions. Here we pr
ovide an extension that uses additional temporal bonds across individual actions
 to enable semantic interpretations of longer videos. Longer temporal connection
s improve scene interpretations as they help discard (temporally) local solution
s in favor of globally superior ones. Using this extension, we demonstrate impro
vements in understanding longer videos, compared to individual interpretations o
f non-overlapping time segments. We verified the success of our approach by gene
rating interpretations for more than 700 video segments from the YouCook data se
t, with intricate videos that exhibit cluttered background, scenarios of occlusi
on, viewpoint variations and changing conditions of illumination. Interpretatio
ns for long video segments were able to yield performance increases of about 70
and, in addition, proved to be more robust to different severe scenarios of clas
sification errors.
****

Line-Sweep: Cross-Ratio For Wide-Baseline Matching and 3D Reconstruction
Srikumar Ramalingam, Michel Antunes, Dan Snow, Gim Hee Lee, Sudeep Pillai; Proce
edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
 2015, pp. 1238-1246
We propose a simple and useful idea based on cross-ratio constraint for wide-bas
eline matching and 3D reconstruction. Most existing methods exploit feature poin
ts and planes from images. Lines have always been considered notorious for both
matching and reconstruction due to the lack of good line descriptors. We propose
 a method to generate and match new points using virtual lines constructed using
 pairs of keypoints, which are obtained using standard feature point detectors.
We use cross-ratio constraints to obtain an initial set of new point matches, wh
ich are subsequently used to obtain line correspondences. We develop a method th
at works for both calibrated and uncalibrated camera configurations. We show com
pelling line-matching and large-scale 3D reconstruction.
****

Simplified Mirror-Based Camera Pose Computation via Rotation Averaging
Gucan Long, Laurent Kneip, Xin Li, Xiaohu Zhang, Qifeng Yu; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1247
-1255
We propose a novel approach to compute the camera pose with respect to a referen
ce object given only mirrored views. The latter originate from a planar mirror a
t different unknown poses. This problem is highly relevant in several extrinsic
camera calibration scenarios, where the camera cannot see the reference object d
irectly. In contrast to numerous existing methods, our approach does not employ
the fixed axis rotation constraint, but represents a more elegant formulation as
 a rotation averaging problem. Our theoretical contribution extends the applicab
ility of rotation averaging to a more general case, and enables mirror-based pos
e estimation in closed-form under the chordal L2-metric, or in an outlier-robust
 way by employing iterative L1-norm averaging. We demonstrate the advantages of
our approach on both synthetic and real data, and show how the method can be app
lied to calibrate the non-overlapping pair of cameras of a common smart phone.
****

On the Relationship Between Visual Attributes and Convolutional Networks
Victor Escorcia, Juan Carlos Niebles, Bernard Ghanem; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1256-1264
One of the cornerstone principles of deep models is their abstraction capacity,
 i.e. their ability to learn abstract concepts from `simpler' ones. Through exte
nsive experiments, we characterize the nature of the relationship between abstra
ct concepts (specifically objects in images) learned by popular and high perform
ing convolutional networks (conv-nets) and established mid-level representations
 used in computer vision (specifically semantic visual attributes). We focus on
attributes due to their impact on several applications, such as object descripti
on, retrieval and mining, and active (and zero-shot) learning. Among the finding
s we uncover, we show empirical evidence of the existence of Attribute Centric
Nodes (ACNs) within a conv-net, which is trained to recognize objects (not attri
butes) in images. These special conv-net nodes (1) collectively encode informati

on pertinent to visual attribute representation and discrimination, (2) are unev enly and sparsely distribution across all layers of the conv-net, and (3) play a n important role in conv-net based object recognition.
****

Saliency Detection by Multi-Context Deep Learning
Rui Zhao, Wanli Ouyang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1265-1274
Low-level saliency cues or priors do not produce good enough saliency detection results especially when the salient object presents in a low-contrast background with confusing visual appearance. This issue raises a serious problem for conve ntional approaches. In this paper, we tackle this problem by proposing a multi-c ontext deep learning framework for salient object detection. We employ deep Conv olutional Neural Networks to model saliency of objects in images. Global context and local context are both taken into account, and are jointly modeled in a uni fied multi-context deep learning framework. To provide a better initialization for training the deep neural networks, we investigate different pre-training st rategies, and a task-specific pre-training scheme is designed to make the multi- context modeling suited for saliency detection. Furthermore, recently proposed c ontemporary deep models in the ImageNet Image Classification Challenge are teste d, and their effectiveness in saliency detection are investigated. Our approach is extensively evaluated on five public datasets, and experimental results show significant and consistent improvements over the state-of-the-art methods.
****

DeepShape: Deep Learned Shape Descriptor for 3D Shape Matching and Retrieval
Jin Xie, Yi Fang, Fan Zhu, Edward Wong; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2015, pp. 1275-1283
Complex geometric structural variations of 3D models usually pose great challeng es in 3D shape matching and retrieval. In this paper, we propose a high-level sh ape feature learning scheme to extract deformation-insensitive feature via a nov el discriminative deep auto-encoder. First, we developed a multiscale shape dist ribution to concisely describe the entire shape of a 3D object. Then, by imposin g the Fisher discrimination criterion on the neurons in the hidden layer, we dev eloped a novel discriminative deep auto-encoder for shape feature learning. Fina lly, the neurons in hidden layers from multiple discriminative auto-encoders are concatenated to form a shape descriptor for 3D shape matching and retrieval. Th e proposed method is evaluated on the representative datasets with large geometr ic variations, i.e., Mcgill, SHREC'10 ShapeGoogle datasets. Experimental results on the benchmark datasets demonstrate the effectiveness of the proposed method on the applications of 3D shape matching and retrieval.
****

Bayesian Adaptive Matrix Factorization With Automatic Model Selection
Peixian Chen, Naiyan Wang, Nevin L. Zhang, Dit-Yan Yeung; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1284-1 292
Low-rank matrix factorization has long been recognized as a fundamental problem in many computer vision applications. Nevertheless, the reliability of existing matrix factorization methods is often hard to guarantee due to challenges brough t by such model selection issues as selecting the noise model and determining th e model capacity. We address these two issues simultaneously in this paper by pr oposing a robust non-parametric Bayesian adaptive matrix factorization (AMF) mod el. AMF proposes a new noise model built on the Dirichlet process Gaussian mixtu re model (DP-GMM) by taking advantage of its high flexibility on component numbe r selection and capability of fitting a wide range of unknown noise. AMF also im poses an automatic relevance determination (ARD) prior on the low-rank factor ma trices so that the rank can be determined automatically without the need for enf orcing any hard constraint. An efficient variational method is then devised for model inference. We compare AMF with state-of-the-art matrix factorization metho ds based on data sets ranging from synthetic data to real-world application data . From the results, AMF consistently achieves better or comparable performance.
****

Joint Action Recognition and Pose Estimation From Video
Bruce Xiaohan Nie, Caiming Xiong, Song-Chun Zhu; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1293-1301
Action recognition and pose estimation from video are closely related tasks for understanding human motion, most methods, however, learn separate models and com bine them sequentially. In this paper, we propose a framework to integrate train ing and testing of the two tasks. A spatial- temporal And-Or graph model is intr oduced to represent action at three scales. Specifically the action is decompose d into poses which are further divided to mid-level ST-parts and then parts. The hierarchical structure of our model captures the geometric and appearance varia tions of pose at each frame and lateral connections between ST-parts at adjacent frames capture the action-specific motion information. The model parameters for three scales are learned discriminatively, and action labels and poses are effi ciently inferred by dynamic programming. Experiments demonstrate that our approa ch achieves state-of-art accuracy in action recognition while also improving pos e estimation.
****

Fast Action Proposals for Human Action Detection and Search
Gang Yu, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1302-1311
In this paper we target at generating generic action proposals in unconstrained videos. Each action proposal corresponds to a temporal series of spatial boundin g boxes, i.e., a spatio-temporal video tube, which has a good potential to locat e one human action. Assuming each action is performed by a human with meaningful motion, both appearance and motion cues are utilized to measure the actionness of the video tubes. After picking those spatiotemporal paths of high actionness scores, our action proposal generation is formulated as a maximum set coverage p roblem, where greedy search is performed to select a set of action proposals tha t can maximize the overall actionness score. Compared with existing action propo sal approaches, our action proposals do not rely on video segmentation and can b e generated in nearly real-time. Experimental results on two challenging dataset s, MSRII and UCF 101, validate the superior performance of our action proposals as well as competitive results on action detection and search.
****

Joint Multi-Feature Spatial Context for Scene Recognition on the Semantic Manifo ld
Xinhang Song, Shuqiang Jiang, Luis Herranz; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1312-1320
In the semantic multinomial framework patches and images are modeled as points i n a semantic probability simplex. Patch theme models are learned resorting to we ak supervision via image labels, which leads the problem of scene categories co-occurring in this semantic space. Fortunately, each category has its own co-occu rrence patterns that are consistent across the images in that category. Thus, di scovering and modeling these patterns is critical to improve the recognition per formance in this representation. In this paper, we observe that not only global co-occurrences at the image-level are important, but also different regions have different category co-occurrence patterns. We exploit local contextual relation s to address the problem of discovering consistent co-occurrence patterns and re moving noisy ones. Our hypothesis is that a less noisy semantic representation, would greatly help the classifier to model consistent co-occurrences and discrim inate better between scene categories. An important advantage of modeling featur es in a semantic space, is that this space is feature independent. Thus, we can combine multiple features and spatial neighbors in the same common space, and fo rmulate the problem as minimizing a context-dependent energy. Experimental resul ts show that exploiting different types of contextual relations consistently imp roves the recognition accuracy. In particular, larger datasets benefit more from the proposed method, leading to very competitive performance.
****

Large-Scale Damage Detection Using Satellite Imagery
Lionel Gueguen, Raffay Hamid; Proceedings of the IEEE Conference on Computer Vis

ion and Pattern Recognition (CVPR), 2015, pp. 1321-1328
Satellite imagery is a valuable source of information for assessing damages in d istressed areas undergoing a calamity, such as an earthquake or an armed conflic t. However, the sheer amount of data required to be inspected for this assessmen t makes it impractical to do it manually. To address this problem, we present a semi-supervised learning framework for large-scale damage detection in satellite imagery. We present a comparative evaluation of our framework using over 88 mil lion images collected from 4,665 square kilometers from 12 different locations a round the world. To enable accurate and efficient damage detection, we introduce a novel use of hierarchical shape features in the bags-of-visual words setting. We analyze how practical factors such as sun, sensor-resolution, and satellite-angle differences impact the effectiveness of our proposed representation, and c ompare it to five alternative features in multiple learning settings. Finally, w e demonstrate through a user-study that our semi-supervised framework results in a ten-fold reduction in human annotation time at a minimal loss in detection ac curacy compared to an exhaustive manual inspection.
****

A Novel Locally Linear KNN Model for Visual Recognition
Qingfeng Liu, Chengjun Liu; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2015, pp. 1329-1337
This paper presents a novel locally linear KNN model with the goal of not only d eveloping efficient representation and classification methods, but also establis hing a relation between them so as to approximate some classification rules, e.g . the Bayes decision rule. Towards that end, first, the proposed model represent s the test sample as a linear combination of all the training samples and derive s a new representation by learning the coefficients considering the reconstructi on, locality and sparsity constraints. The theoretical analysis shows that the new representation has the grouping effect of the nearest neighbors, which is ab le to approximate the "ideal representation". And then the locally linear KNN mo del based classifier (LLKNNC), which shows its connection to the Bayes decision rule for minimum error in the view of kernel density estimation, is proposed for classification. Besides, the locally linear nearest mean classifier (LLNMC), wh ose relation to the LLKNNC is just like the nearest mean classifier to the KNN c lassifier, is also derived. Furthermore, to provide reliable kernel density esti mation, the shifted power transformation and the coefficients cut-off method are applied to improve the performance of the proposed method. The effectiveness of the proposed model is evaluated on several visual recognition tasks such as fac e recognition, scene recognition, object recognition and action recognition. The experimental results show that the proposed model is effective and outperforms some other representative popular methods.
****

Bilinear Random Projections for Locality-Sensitive Binary Codes
Saehoon Kim, Seungjin Choi; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2015, pp. 1338-1346
Locality-sensitive hashing (LSH) is a popular data-independent indexing method f or approximate similarity search, where random projections followed by quantizat ion hash the points from the database so as to ensure that the probability of co llision is much higher for objects that are close to each other than for those t hat are far apart. Most of high-dimensional visual descriptors for images exhibi t a natural matrix structure. When visual descriptors are represented by high-di mensional feature vectors and long binary codes are assigned, a random projectio n matrix requires expensive complexities in both space and time. In this paper w e analyze a bilinear random projection method where feature matrices are transfo rmed to binary codes by two smaller random projection matrices. We base our theo retical analysis on extending Raginsky and Lazebnik's result where random Fourie r features are composed with random binary quantizers to form locality sensitive binary codes. To this end, we answer the following two questions: (1) whether a bilinear random projection also yields similarity-preserving binary codes; (2) whether a bilinear random projection yields performance gain or loss, compared t o a large linear projection. Regarding the first question, we present upper and

lower bounds on the expected Hamming distance between binary codes produced by b
ilinear random projections. In regards to the second question, we analyze the up
per and lower bounds on covariance between two bits of binary codes, showing tha
t the correlation between two bits is small. Numerical experiments on MNIST and
Flickr45K datasets confirm the validity of our method.
****

Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks f
or Human Pose Estimation
Xiaochuan Fan, Kang Zheng, Yuewei Lin, Song Wang; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1347-1355
We propose a new learning-based method for estimating 2D human pose from a singl
e image, using Dual-Source Deep Convolutional Neural Networks (DS-CNN). Recently
, many methods have been developed to estimate human pose by using pose priors t
hat are estimated from physiologically inspired graphical models or learned from
 a holistic perspective. In this paper, we propose to integrate both the local (
body) part appearance and the holistic view of each local part for more accurate
 human pose estimation. Specifically, the proposed DS-CNN takes a set of image p
atches (category-independent object proposals for training and multi-scale slidi
ng windows for testing) as the input and then learns the appearance of each loca
l part by considering their holistic views in the full body. Using DS-CNN, we ac
hieve both joint detection, which determines whether an image patch contains a b
ody joint, and joint localization, which finds the exact location of the joint i
n the image patch. Finally, we develop an algorithm to combine these joint detec
tion/localization results from all the image patches for estimating the human po
se. The experimental results show the effectiveness of the proposed method by co
mparing to the state-of-the-art human-pose estimation methods based on pose prio
rs that are estimated from physiologically inspired graphical models or learned
from a holistic perspective.
****

Superpixel Segmentation Using Linear Spectral Clustering
Zhengqin Li, Jiansheng Chen; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2015, pp. 1356-1363
We present in this paper a superpixel segmentation algorithm called Linear Spect
ral Clustering (LSC), which produces compact and uniform superpixels with low co
mputational costs. Basically, a normalized cuts formulation of the superpixel se
gmentation is adopted based on a similarity metric that measures the color simil
arity and space proximity between image pixels. However, instead of using the tr
aditional eigen-based algorithm, we approximate the similarity metric using a ke
rnel function leading to an explicitly mapping of pixel values and coordinates i
nto a high dimensional feature space. We revisit the conclusion that by appropri
ately weighting each point in this feature space, the objective functions of wei
ghted K-means and normalized cuts share the same optimum point. As such, it is p
ossible to optimize the cost function of normalized cuts by iteratively applying
 simple K-means clustering in the proposed feature space. LSC is of linear compu
tational complexity and high memory efficiency and is able to preserve global pr
operties of images. Experimental results show that LSC performs equally well or
better than state of the art superpixel segmentation algorithms in terms of seve
ral commonly used evaluation metrics in image segmentation.
****

Person Count Localization in Videos From Noisy Foreground and Detections
Sheng Chen, Alan Fern, Sinisa Todorovic; Proceedings of the IEEE Conference on C
omputer Vision and Pattern Recognition (CVPR), 2015, pp. 1364-1372
This paper formulates and presents a solution to a new problem called person cou
nt localization. Given a video of a crowded scene, our goal is to output for eac
h frame a set of: 1) Detections optimally covering both isolated individuals and
 cluttered groups of people; and 2) Counts of people inside these detections. Th
is problem is a middle-ground between frame-level person counting, which does no
t localize counts, and person detection aimed at perfectly localizing people wit
h count-one detections. Our problem formulation is important for a wide range o
f domains, where people appear frequently under severe occlusion within a crowd.

As these crowds are often visually distinct from the rest of the scene, they ca n be viewed as ``visual phrases'' whose spatially tight localization and count a ssignment could facilitate higher-level video understanding. For count localizat ion, we specify a novel framework of iterative error-driven revisions of a flow graph derived from noisy input of people detections and foreground segmentation. Each iteration creates and solves an integer program for count localization bas ed on iterative revisions of the flow graph. The graph revisions are based on de tected violations of basic integrity constraints. They in turn trigger learned m odifications to the graph aimed at reducing noise in input features. For evaluat ion, we introduce a new metric that measures both count precision and localizati on of our approach on American football and pedestrian videos.
****

Good Features to Track for Visual SLAM
Guangcong Zhang, Patricio A. Vela; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2015, pp. 1373-1382
Not all measured features in SLAM/SfM contribute to accurate localization during the estimation process, thus it is sensible to utilize only those that do. This paper describes a method for selecting a subset of features that are of high ut ility for localization in the SLAM/SfM estimation process. It is derived by exa mining the observability of SLAM and, being complimentary to the estimation proc ess, it easily integrates into existing SLAM systems. The measure of estimation utility is formulated with temporal and instantaneous observability indices. Ef ficient computation strategies for the observability indices are described based on incremental singular value decomposition (SVD) and greedy selection for the temporal and instantaneous observability indices, respectively. The greedy sele ction is near-optimal since the observability index is (approximately) submodula r. The proposed method improves localization and data association. Controlled s ynthetic experiments with ground truth demonstrate the improved localization acc uracy, and real-time SLAM experiments demonstrate the improved data association.
****

Discovering States and Transformations in Image Collections
Phillip Isola, Joseph J. Lim, Edward H. Adelson; Proceedings of the IEEE Confere nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1383-1391
Objects in visual scenes come in a rich variety of transformed states. A few cla sses of transformation have been heavily studied in computer vision: mostly simp le, parametric changes in color and geometry. However, transformations in the ph ysical world occur in many more flavors, and they come with semantic meaning: e. g., bending, folding, aging, etc. The transformations an object can undergo tell us about its physical and functional properties. In this paper, we introduce a dataset of objects, scenes, and materials, each of which is found in a variety o f transformed states. Given a novel collection of images, we show how to explain the collection in terms of the states and transformations it depicts. Our syste m works by generalizing across object classes: states and transformations learne d on one set of objects are used to interpret the image collection for an entire ly new object class.
****

Generalized Deformable Spatial Pyramid: Geometry-Preserving Dense Correspondence Estimation
Junhwa Hur, Hwasup Lim, Changsoo Park, Sang Chul Ahn; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1392-1400
We present a Generalized Deformable Spatial Pyramid (GDSP) matching algorithm fo r calculating the dense correspondence between a pair of images with large appea rance variations. The main challenges of the problem generally originate in appe arance dissimilarities and geometric variations between images. To address these challenges, we improve the existing Deformable Spatial Pyramid (DSP) model by g eneralizing the search space and devising the spatial smoothness. The former is leveraged by rotations and scales, and the latter simultaneously considers depen dencies between high-dimensional labels through the pyramid structure. Our spati al regularization in the high-dimensional space enables our model to effectively preserve the meaningful geometry of objects in the input images while allowing

for a wide range of geometry variations such as perspective transform and non-rigid deformation. The experimental results on public datasets and challenging scenarios show that our method outperforms the state-of-the-art methods both qualitatively and quantitatively.
****

Classifier Adaptation at Prediction Time
Amelie Royer, Christoph H. Lampert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1401-1409
Classifiers for object categorization are usually evaluated by their accuracy on a set of i.i.d. test examples. This provides us with an estimate of the expected error when applying the classifiers to a single new image. In real application, however, classifiers are rarely only used for a single image and then discarded. Instead, they are applied sequentially to many images, and these are typically not i.i.d. samples from a fixed data distribution, but they carry dependencies and their class distribution varies over time. In this work, we argue that the phenomenon of correlated data at prediction time is not a nuisance, but a blessing in disguise. We describe a probabilistic method for adapting classifiers at prediction time without having to retrain them. We also introduce a framework for creating realistically distributed image sequences, which offers a way to benchmark classifier adaptation methods, such as the one we propose. Experiments on the ILSVRC2010 and ILSVRC2012 datasets show that adapting object classification systems at prediction time can significantly reduce their error rate, even with no additional human feedback.
****

Phase-Based Frame Interpolation for Video
Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, Alexander Sorkine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1410-1418
Standard approaches to computing interpolated (in-between) frames in a video sequence require accurate pixel correspondences between images e.g. using optical flow. We present an efficient alternative by leveraging recent developments in phase-based methods that represent motion in the phase shift of individual pixels. This concept allows in-between images to be generated by simple per-pixel phase modification, without the need for any form of explicit correspondence estimation. Up until now, such methods have been limited in the range of motion that can be interpolated, which fundamentally restricts their usefulness. In order to reduce these limitations, we introduce a novel, bounded phase shift correction method that combines phase information across the levels of a multi-scale pyramid. Additionally, we propose extensions for phase-based image synthesis that yield smoother transitions between the interpolated images. Our approach avoids expensive global optimization typical of optical flow methods, and is both simple to implement and easy to parallelize. This allows us to interpolate frames at a fraction of the computational cost of traditional optical flow-based solutions, while achieving similar quality and in some cases even superior results. Our method fails gracefully in difficult interpolation settings, e.g., significant appearance changes, where flow-based methods often introduce serious visual artifacts. Due to its efficiency, our method is especially well suited for frame interpolation and retiming of high resolution, high frame rate video.
****

Matching-CNN Meets KNN: Quasi-Parametric Human Parsing
Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1419-1427
Both parametric and non-parametric approaches have demonstrated encouraging performances in the human parsing task, namely segmenting a human image into several semantic regions (e.g., hat, bag, left arm, face). In this work, we aim to develop a new solution with the advantages of both methodologies, namely supervision from annotated data and the flexibility to use newly annotated (possibly uncommon) images, and present a quasi-parametric human parsing model. Under the classic KNN-based nonparametric framework, the parametric Matching Convolutional Neur

al Network (M-CNN) is proposed to predict the matching confidence and displaceme nt of the best matched region in the testing image for a particular semantic reg ion in one KNN image. Given a testing image, we first retrieve its KNN images fr om the annotated/manually-parsed human image corpus. Then each semantic region i n each KNN image is matched with confidence to the testing image using M-CNN, an d the matched regions from all KNN images are further fused, followed by a super pixel smoothing procedure to obtain the ultimate human parsing result. The M-CNN differs from the classic CNN in that the tailored cross image matching filters are introduced to characterize the matching between the testing image and the se mantic region of a KNN image. The cross image matching filters are defined at di fferent convolution layers, each aiming to capture a particular range of displac ements. Comprehensive evaluations over a large dataset with 7,700 annotated huma n images well demonstrate the significant performance gain from the quasi-parame tric model over the state-of-the-arts, for the human parsing task.

****

Absolute Pose for Cameras Under Flat Refractive Interfaces

Sebastian Haner, Kalle Astrom; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2015, pp. 1428-1436

This paper studies the problem of determining the absolute pose of a perspective camera observing a scene through a known refractive plane, the flat boundary be tween transparent media with different refractive indices. Efficient minimal sol vers are developed for the 2D, known orientation and known rotation axis cases, and near-minimal solvers for the general calibrated and unknown focal length cas es. We show that ambiguities in the equations of Snell's law give rise to a larg e number of false solutions, increasing the complexity of the problem. Evaluatio n of the solvers on both synthetic and real data show excellent numerical perfor mance, and the necessity of explicitly modelling refraction to obtain accurate p ose estimates.

****

Protecting Against Screenshots: An Image Processing Approach

Alex Yong-Sang Chia, Udana Bandara, Xiangyu Wang, Hiromi Hirano; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1437-1445

Motivated by reasons related to data security and privacy, we propose a method t o limit meaningful visual contents of a display from being captured by screensho ts. Traditional methods take a system architectural approach to protect against screenshots. We depart from this framework, and instead exploit image processing techniques to distort visual data of a display and present the distorted data t o the viewer. Given that a screenshot captures distorted visual contents, it yie lds limited useful data. We exploit the human visual system to empower viewers t o automatically and mentally recover the distorted contents into a meaningful fo rm in real-time. Towards this end, we leverage on findings from psychological st udies which show that blending of visual information from recent and current fix ations enables human to form meaningful representation of a scene. We model this blending of information by an additive process, and exploit this to design a vi sual contents distortion algorithm that supports real-time contents recovery by the human visual system. Our experiments and user study demonstrate the feasibil ity of our method to allow viewers to readily interpret visual contents of a dis play, while limiting meaningful contents from being captured by screenshots.

****

Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction

Ijaz Akhter, Michael J. Black; Proceedings of the IEEE Conference on Computer Vi sion and Pattern Recognition (CVPR), 2015, pp. 1446-1455

Estimating 3D human pose from 2D joint locations is central to the analysis of p eople in images and video. To address the fact that the problem is inherently il l posed, many methods impose a prior over human poses. Unfortunately these prior s admit invalid poses because they do not model how joint-limits vary with pose. Here we make two key contributions. First, we collect a motion capture dataset that explores a wide range of human poses. From this we learn a pose-dependent m odel of joint limits that forms our prior. Both dataset and prior are available

for research purposes. Second, we define a general parametrization of body pose and a new, multi-stage, method to estimate 3D pose from 2D joint locations using an over-complete dictionary of poses. Our method shows good generalization while avoiding impossible poses. We quantitatively compare our method with recent work and show state-of-the-art results on 2D to 3D pose estimation using the CMU mocap dataset. We also show superior results using manual annotations on real images and automatic detections on the Leeds sports pose dataset.
****

VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases

Fereshteh Sadeghi, Santosh K. Kumar Divvala, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1456-1464

How can we know whether a statement about our world is valid. For example, given a relationship between a pair of entities e.g., `eat(horse, hay)', how can we know whether this relationship is true or false in general. Gathering such knowledge about entities and their relationships is one of the fundamental challenges in knowledge extraction. Most previous works on knowledge extraction have focused purely on text-driven reasoning for verifying relation phrases. In this work, we introduce the problem of visual verification of relation phrases and developed a Visual Knowledge Extraction system called VisKE. Given a verb-based relation phrase between common nouns, our approach assess its validity by jointly analyzing over text and images and reasoning about the spatial consistency of the relative configurations of the entities and the relation involved. Our approach involves no explicit human supervision thereby enabling large-scale analysis. Using our approach, we have already verified over 12000 relation phrases. Our approach has been used to not only enrich existing textual knowledge bases by improving their recall, but also augment open-domain question-answer reasoning.
****

A Graphical Model Approach for Matching Partial Signatures

Xianzhi Du, David Doermann, Wael Abd-Almageed; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1465-1472

In this paper, we present a novel partial signature matching method using graphical models. Shape context features are extracted from the contour of signatures to capture local variations, and K-means clustering is used to build a visual vocabulary from a set of reference signatures. To describe the signatures, supervised latent Dirichlet allocation is used to learn the latent distributions of the salient regions over the visual vocabulary and hierarchical Dirichlet processes are implemented to infer the number of salient regions needed. Our work is evaluated on three datasets derived from the DS-I Tobacco signature dataset with clean signatures and the DS-II UMD dataset with signatures with different degradations. The results show the effectiveness of the approach for both the partial and full signature matching.
****

From Captions to Visual Concepts and Back

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1473-1482

This paper presents a novel approach for automatically generating image descriptions: visual detectors, language models, and multimodal similarity models learnt directly from a dataset of image captions. We use multiple instance learning to train visual detectors for words that commonly occur in captions, including many different parts of speech such as nouns, verbs, and adjectives. The word detector outputs serve as conditional inputs to a maximum-entropy language model. The language model learns from a set of over 400,000 image descriptions to capture the statistics of word usage. We capture global semantics by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model. Our system is state-of-the-art on the official Microsoft COCO benchmark, producing a BLEU-4 score of 29.1%. When human judges compare the system captions to on

es written by other people on our held-out test set, the system captions have eq
ual or better quality 34% of the time.
****

Semi-Supervised Low-Rank Mapping Learning for Multi-Label Classification
Liping Jing, Liu Yang, Jian Yu, Michael K. Ng; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1483-1491
Multi-label problems arise in various domains including automatic multimedia dat
a categorization, and have generated significant interest in computer vision and
 machine learning community. However, existing methods do not adequately addres
s two key challenges: exploiting correlations between labels and making up for t
he lack of labeled data or even missing labels. In this paper, we proposed a sem
i-supervised low-rank mapping (SLRM) model to handle these two challenges. SLRM
model takes advantage of the nuclear norm regularization on mapping to effective
ly capture the label correlations. Meanwhile, it introduces manifold regularizer
 on mapping to capture the intrinsic structure among data, which provides a good
 way to reduce the required labeled data with improving the classification perfo
rmance. Furthermore, we designed an efficient algorithm to solve SLRM model base
d on alternating direction method of multipliers and thus it can efficiently dea
l with large-scale data sets. Experiments on four real-world multimedia data set
s demonstrate that the proposed method can exploit the label correlations and ob
tain promising and better label prediction results than state-of-the-art methods
.
****

ConceptLearner: Discovering Visual Concepts From Weakly Labeled Image Collection
s
Bolei Zhou, Vignesh Jagadeesh, Robinson Piramuthu; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1492-1500
Discovering visual knowledge from weakly labeled data is crucial to scale up com
puter vision recognition systems, since it is expensive to obtain fully labeled
data for a large number of concept categories. In this paper, we propose Concept
Learner, which is a scalable approach to discover visual concepts from weakly la
beled image collections. Thousands of visual concept detectors are learned autom
atically, without human in the loop for additional annotation. We show that thes
e learned detectors could be applied to recognize concepts at image-level and to
 detect concepts at image region-level accurately. Under domain-specific supervi
sion, we further evaluate the learned concepts for scene recognition on SUN data
base and for object detection on Pascal VOC 2007. ConceptLearner shows promising
 performance compared to fully supervised and weakly supervised methods.
****

Computationally Bounded Retrieval
Mohammad Rastegari, Cem Keskin, Pushmeet Kohli, Shahram Izadi; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1
501-1509
The increase in size of large image databases makes the problem of efficient ret
rieval extremely challenging. This is especially true in the case of high dimens
ional data where even operations like hashing become expensive because of costly
 projection operators. Unlike most hashing methods that sacrifice accuracy for s
peed, we propose a novel method that improves the speed of high dimensional imag
e retrieval by several orders of magnitude without any significant drop in perfo
rmance. To do this, we propose to learn computationally bounded sparse projectio
ns for the encoding step. To further increase the accuracy of the method, we add
 an orthogonality constraint on projections to reduce bit correlation. We then i
ntroduce an iterative scheme that jointly optimizes this objective, which helps
us obtain fast and efficient projections. We demonstrate this technique on large
 retrieval databases, specifically ImageNET, GIST1M and SUN-attribute for the ta
sk of nearest neighbor retrieval, and show that our method achieves a speed-up o
f up to a factor of 100 over state-of-the-art methods, while having on-par and i
n some cases even better accuracy.
****

Viewpoints and Keypoints

Shubham Tulsiani, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1510-1519

We characterize the problem of pose estimation for rigid objects in terms of determining viewpoint to explain coarse pose and keypoint prediction to capture the finer details. We address both these tasks in two different settings - the constrained setting with known bounding boxes and the more challenging detection setting where the aim is to simultaneously detect and correctly estimate pose of objects. We present Convolutional Neural Network based architectures for these and demonstrate that leveraging viewpoint estimates can substantially improve local appearance based keypoint predictions. In addition to achieving significant improvements over state-of-the-art in the above tasks, we analyze the error modes and effect of object characteristics on performance to guide future efforts towards this goal.

****

Discrete Hyper-Graph Matching

Junchi Yan, Chao Zhang, Hongyuan Zha, Wei Liu, Xiaokang Yang, Stephen M. Chu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1520-1528

This paper focuses on the problem of hyper-graph matching, by accounting for both unary and higher-order affinity terms. Our method is in line with the linear approximate framework while the problem is iteratively solved in discrete space. It is empirically found more efficient than many extant continuous methods. Moreover, it avoids unknown accuracy loss by heuristic rounding step from the continuous approaches. Under weak assumptions, we prove the iterative discrete gradient assignment in general will trap into a degenerating case -- an m-circle solution path where m is the order of the problem. A tailored adaptive relaxation mechanism is devised to detect the degenerating case and makes the algorithm converge to a fixed point in discrete space. Evaluations on both synthetic and real-world data corroborate the efficiency of our method.

****

Rolling Shutter Motion Deblurring

Shuochen Su, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1529-1537

Although motion blur and rolling shutter deformations are closely coupled artifacts in images taken with CMOS image sensors, the two phenomena have so far mostly been treated separately, with deblurring algorithms being unable to handle rolling shutter wobble, and rolling shutter algorithms being incapable of dealing with motion blur. We propose an approach that delivers sharp and undistorted output given a single rolling shutter motion blurred image. The key to achieving this is a global modeling of the camera motion trajectory, which enables each scan line of the image to be deblurred with the corresponding motion segment. We show the results of the proposed framework through experiments on synthetic and real data.

****

Learning to Generate Chairs With Convolutional Neural Networks

Alexey Dosovitskiy, Jost Tobias Springenberg, Thomas Brox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1538-1546

We train a generative convolutional neural network which is able to generate images of objects given object type, viewpoint, and color. We train the network in a supervised manner on a dataset of rendered 3D chair models. Our experiments show that the network does not merely learn all images by heart, but rather finds a meaningful representation of a 3D chair model allowing it to assess the similarity of different chairs, interpolate between given viewpoints to generate the missing ones, or invent new chair styles by interpolating between chairs from the training set. We show that the network can be used to find correspondences between different chairs from the dataset, outperforming existing approaches on this task.

****

Accurate Depth Map Estimation From a Lenslet Light Field Camera

Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1547-1555

This paper introduces an algorithm that accurately estimates depth maps using a lenslet light field camera. The proposed algorithm estimates the multi-view stereo correspondences with sub-pixel accuracy using the cost volume. The foundation for constructing accurate costs is threefold. First, the sub-aperture images are displaced using the phase shift theorem. Second, the gradient costs are adaptively aggregated using the angular coordinates of the light field. Third, the feature correspondences between the sub-aperture images are used as additional constraints. With the cost volume, the multi-label optimization propagates and corrects the depth map in the weak texture regions. Finally, the local depth map is iteratively refined through fitting the local quadratic function to estimate a non-discrete depth map. Because micro-lens images contain unexpected distortions, a method is also proposed that corrects this error. The effectiveness of the proposed algorithm is demonstrated through challenging real world examples and including comparisons with the performance of advanced depth estimation algorithms.

****

Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval

Fang Zhao, Yongzhen Huang, Liang Wang, Tieniu Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1556-1564

With the rapid growth of web images, hashing has received increasing interests in large scale image retrieval. Research efforts have been devoted to learning compact binary codes that preserve semantic similarity based on labels. However, most of these hashing methods are designed to handle simple binary similarity. The complex multilevel semantic structure of images associated with multiple labels have not yet been well explored. Here we propose a deep semantic ranking based method for learning hash functions that preserve multilevel semantic similarity between multi-label images. In our approach, deep convolutional neural network is incorporated into hash functions to jointly learn feature representations and mappings from them to hash codes, which avoids the limitation of semantic representation power of hand-crafted features. Meanwhile, a ranking list that encodes the multilevel similarity information is employed to guide the learning of such deep hash functions. An effective scheme based on surrogate loss is used to solve the intractable optimization problem of nonsmooth and multivariate ranking measures involved in the learning procedure. Experimental results show the superiority of our proposed approach over several state-of-the-art hashing methods in term of ranking evaluation metrics when tested on multi-label image datasets.

****

Similarity Learning on an Explicit Polynomial Kernel Feature Map for Person Re-Identification

Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, Jingdong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1565-1573

In this paper, we address the person re-identification problem, discovering the correct matches for a probe person image from a set of gallery person images. We follow the learning-to-rank methodology and learn a similarity function to maximize the difference between the similarity scores of matched and unmatched images for a same person. We introduce at least three contributions to person re-identification. First, we present an explicit polynomial kernel feature map, which is capable of characterizing the similarity information of all pairs of patches between two images, called soft-patch-match, instead of greedily keeping only the best matched patch, and thus more robust. Second, we introduce a mixture of linear similarity functions that is able to discover different soft-patch-matching patterns. Last, we introduce a negative semi-definite regularization over a subset of the weights in the similarity function, which is motivated by the connection between explicit polynomial kernel feature map and the Mahalanobis distance, as well as the sparsity constraint over the parameters to avoid over-fitting. Experimental results over three public benchmarks demonstrate the superiority of our approach.

****
Learning to Propose Objects
Philipp Krahenbuhl, Vladlen Koltun; Proceedings of the IEEE Conference on Comput er Vision and Pattern Recognition (CVPR), 2015, pp. 1574-1582
We present an approach for highly accurate bottom-up object segmentation. Given an image, the approach rapidly generates a set of regions that delineate candida te objects in the image. The key idea is to train an ensemble of figure-ground s egmentation models. The ensemble is trained jointly, enabling individual models to specialize and complement each other. We reduce ensemble training to a sequen ce of uncapacitated facility location problems and show that highly accurate seg mentation ensembles can be trained by combinatorial optimization. The training p rocedure jointly optimizes the size of the ensemble, its composition, and the pa rameters of incorporated models, all for the same objective. The ensembles opera te on elementary image features, enabling rapid image analysis. Extensive experi ments demonstrate that the presented approach outperforms prior object proposal algorithms by a significant margin, while having the lowest running time. The tr ained ensembles generalize across datasets, indicating that the presented approa ch is capable of learning a generally applicable model of bottom-up segmentation .
****
Basis Mapping Based Boosting for Object Detection
Haoyu Ren, Ze-Nian Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1583-1591
We propose a novel mapping method to improve the training accuracy and efficien cy of boosted classifiers for object detection. The key step of the proposed met hod is a non linear mapping on original samples by referring to the basis sample s before feeding into the weak classifiers, where the basis samples correspond t o the hard samples in the current training stage. We show that the basis mapping based weak classifier is an approximation of kernel weak classifiers while keep ing the same computation cost as linear weak classifiers. As a result, boosting with such weak classifiers is more effective. In this paper, two different non-l inear mappings are shown to work well. We adopt the LogitBoost algorithm to trai n the weak classifiers based on the Histogram of Oriented Gradient descriptor (HOG). Experimental results show that the proposed approach significantly improv es the detection accuracy and training efficiency of the boosted classifier. It also achieves performance comparable with the commonly used methods on public da tasets for both pedestrian detection and general object detection tasks.
****
Computing the Stereo Matching Cost With a Convolutional Neural Network
Jure Zbontar, Yann LeCun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1592-1599
We present a method for extracting depth information from a rectified image pair . We train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. The cost is refined by cro ss-based cost aggregation and semiglobal matching, followed by a left-right cons istency check to eliminate errors in the occluded regions. Our stereo method ach ieves an error rate of 2.61 % on the KITTI stereo dataset and is currently (Augu st 2014) the top performing method on this dataset.
****
Recognize Complex Events From Static Images by Fusing Deep Channels
Yuanjun Xiong, Kai Zhu, Dahua Lin, Xiaoou Tang; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1600-1609
A considerable portion of web images capture events that occur in our personal l ives or social activities. In this paper, we aim to develop an effective method for recognizing events from such images. Despite the sheer amount of study on ev ent recognition, most existing methods rely on videos and are not directly appli cable to this task. Generally, events are complex phenomena that involve interac tions among people and objects, and therefore analysis of event photos requires techniques that can go beyond recognizing individual objects and carry out joint reasoning based on evidences of multiple aspects. Inspired by the recent succes

s of deep learning, we formulate a multi-layer framework to tackle this problem, which takes into account both visual appearance and the interactions among humans and objects, and combines them via semantic fusion. An important issue arising here is that humans and objects discovered by detectors are in the form of bounding boxes, and there is no straightforward way to represent their interactions and incorporate them with a deep network. We address this using a novel strategy that projects the detected instances onto multi-scale spatial maps. On a large dataset with $60,000$ images, the proposed method achieved substantial improvement over the state-of-the-art, raising the accuracy of event recognition by over 10%.

****

Multi-Feature Max-Margin Hierarchical Bayesian Model for Action Recognition
Shuang Yang, Chunfeng Yuan, Baoxin Wu, Weiming Hu, Fangshi Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1610-1618

In this paper, a multi-feature max-margin hierarchical Bayesian model (M3HBM) is proposed for action recognition. Different from existing methods which separate representation and classification into two steps, M3HBM jointly learns a high-level representation by combining a hierarchical generative model (HGM) and discriminative max-margin classifiers in a unified Bayesian framework. Specifically, HGM is proposed to represent actions by distributions over latent spatial temporal patterns (STPs) which are learned from multiple feature modalities and shared among different classes. For recognition, we employ Gibbs classifiers to minimize the expected loss function based on the max-margin principle and use the classifiers as regularization terms of M3HBM to perform Bayeisan estimation for classifier parameters together with the learning of STPs. In addition, multi-task learning is applied to learn the model from multiple feature modalities for different classes. For test videos, we obtain the representations by the inference process and perform action recognition by the learned Gibbs classifiers. For the learning and inference process, we derive an efficient Gibbs sampling algorithm to solve the proposed M3HBM. Extensive experiments on several datasets demonstrate both the representation power and the classification capability of our approach for action recognition.

****

Model Recommendation: Generating Object Detectors From Few Samples
Yu-Xiong Wang, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1619-1628

In this paper, we explore an approach to generating detectors that is radically different from the conventional way of learning a detector from a large corpus of annotated positive and negative data samples. Instead, we assume that we have evaluated ''off-line'' a large library of detectors against a large set of detection tasks. Given a new target task, we evaluate a subset of the models on few samples from the new task and we use the matrix of models-tasks ratings to predict the performance of all the models in the library on the new task, enabling us to select a good set of detectors for the new task. This approach has three key advantages of great interest in practice: 1) generating a large collection of expressive models in an unsupervised manner is possible; 2) a far smaller set of annotated samples is needed compared to that required for training from scratch; and 3) recommending models is a very fast operation compared to the notoriously expensive training procedures of modern detectors. (1) will make the models informative across different categories; (2) will dramatically reduce the need for manually annotating vast datasets for training detectors; and (3) will enable rapid generation of new detectors.

****

A Linear Least-Squares Solution to Elastic Shape-From-Template
Abed Malti, Adrien Bartoli, Richard Hartley; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1629-1637

We cast SfT (Shape-from-Template) as the search of a vector field (X,dX), composed of the pose X and the displacement dX that produces the deformation. We propose the first fully linear least-squares SfT method modeling elastic deformation

s. It relies on a set of Solid Boundary Constraints SBC to position the template at X in the deformed frame. The displacement is mapped by the stiffness matrix to minimize the amount of force responsible for the deformation. This linear minimization is subjected to the Reprojection Boundary Constraints RBC of the deformed shape X+dX on the deformed image. Compared to state-of-the-art methods, this new formulation allows us to obtain accurate results at a low computational cost.

****

Robust Large Scale Monocular Visual SLAM

Guillaume Bourmaud, Remi Megret; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1638-1647

This paper deals with the trajectory estimation of a monocular calibrated camera evolving in a large unknown environment, also known as monocular visual simultaneous localization and mapping. The contribution of this paper is threefold: 1) We develop a new formalism that builds upon the so called Known Rotation Problem to robustly estimate submaps (parts of the camera trajectory and the unknown environment). 2) In order to obtain a globally consistent map (up to a scale factor), we propose a novel loopy belief propagation algorithm that is able to efficiently align a large number of submaps. Our approach builds a graph of relative 3D similarities (computed between the submaps) and estimates the global 3D similarities by passing messages through a super graph until convergence. 3) To render the whole framework more robust, we also propose a simple and efficient outlier removal algorithm that detects outliers in the graph of relative 3D similarities. We extensively demonstrate, on the TUM and KITTI benchmarks as well as on several other challenging video sequences, that the proposed method outperforms the state of the art algorithms.

****

Membership Representation for Detecting Block-Diagonal Structure in Low-Rank or Sparse Subspace Clustering

Minsik Lee, Jieun Lee, Hyeogjin Lee, Nojun Kwak; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1648-1656

Recently, there have been many proposals with state-of-the-art results in subspace clustering that take advantages of the low-rank or sparse optimization techniques. These methods are based on self-expressive models, which have well-defined theoretical aspects. They produce matrices with (approximately) block-diagonal structure, which is then applied to spectral clustering. However, there is no definitive way to construct affinity matrices from these block-diagonal matrices and it is ambiguous how the performance will be affected by the construction method. In this paper, we propose an alternative approach to detect block-diagonal structures from these matrices. The proposed method shares the philosophy of the above subspace clustering methods, in that it is a self-expressive system based on a Hadamard product of a membership matrix. To resolve the difficulty in handling the membership matrix, we solve the convex relaxation of the problem and then transform the representation to a doubly stochastic matrix, which is closely related to spectral clustering. The result of our method has eigenvalues normalized in between zero and one, which is more reliable to estimate the number of clusters and to perform spectral clustering. The proposed method shows competitive results in our experiments, even though we simply count the number of eigenvalues larger than a certain threshold to find the number of clusters.

****

Bayesian Inference for Neighborhood Filters With Application in Denoising

Chao-Tsung Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1657-1665

Range-weighted neighborhood filters are useful and popular for their edge-preserving property and simplicity, but they are originally proposed as intuitive tools. Previous works needed to connect them to other tools or models for indirect property reasoning or parameter estimation. In this paper, we introduce a unified empirical Bayesian framework to do both directly. A neighborhood noise model is proposed to reason and infer the Yaroslavsky, bilateral, and modified non-local means filters. An EM+ algorithm is devised to estimate the essential parameter,

range variance, via the model fitting to empirical distributions. Finally, we apply this framework to color-image denoising. Experimental results show that the proposed model fits noisy images well and the range variance is estimated successfully. The image quality can also be improved by a proposed recursive fitting and filtering scheme.
****

Deep LAC: Deep Localization, Alignment and Classification for Fine-Grained Recognition
Di Lin, Xiaoyong Shen, Cewu Lu, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1666-1674
We propose a fine-grained recognition system that incorporates part localization, alignment, and classification in one deep neural network. This is a nontrivial process, as the input to the classification module should be functions that enable back-propagation in constructing the solver. Our major contribution is to propose a valve linkage function(VLF) for back-propagation chaining and form our deep localization, alignment and classification (LAC) system. The VLF can adaptively compromise the errors of classification and alignment when training the LAC model. It in turn helps update localization. The performance on fine-grained object data bears out the effectiveness of our LAC system.
****

Unconstrained Realtime Facial Performance Capture
Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, Hao Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1675-1683
We introduce a realtime facial tracking system specifically designed for performance capture in unconstrained settings using a consumer-level RGB-D sensor. Our framework provides uninterrupted 3D facial tracking, even in the presence of extreme occlusions such as those caused by hair, hand-to-face gestures, and wearable accessories. Anyone's face can be instantly tracked and the users can be switched without an extra calibration step. During tracking, we explicitly segment face regions from any occluding parts by detecting outliers in the shape and appearance input using an exponentially smoothed and user-adaptive tracking model as prior. Our face segmentation combines depth and RGB input data and is also robust against illumination changes. To enable continuous and reliable facial feature tracking in the color channels, we synthesize plausible face textures in the occluded regions. Our tracking model is personalized on-the-fly by progressively refining the user's identity, expressions, and texture with reliable samples and temporal filtering. We demonstrate robust and high-fidelity facial tracking on a wide range of subjects with highly incomplete and largely occluded data. Our system works in everyday environments and is fully unobtrusive to the user, impacting consumer AR applications and surveillance.
****

Blind Optical Aberration Correction by Exploring Geometric and Visual Priors
Tao Yue, Jinli Suo, Jue Wang, Xun Cao, Qionghai Dai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1684-1692
Optical aberration widely exists in optical imaging systems, especially in consumer-level cameras. In contrast to previous solutions using hardware compensation or pre-calibration, we propose a computational approach for blind aberration removal from a single image, by exploring various geometric and visual priors. The global rotational symmetry allows us to transform the non-uniform degeneration into several uniform ones by the proposed radial splitting and warping technique. Locally, two types of symmetry constraints, i.e. central symmetry and reflection symmetry are defined as geometric priors in central and surrounding regions, respectively. Furthermore, by investigating the visual artifacts of aberration degenerated images captured by consumer-level cameras, the non-uniform distribution of sharpness across color channels and the image lattice is exploited as visual priors, resulting in a novel strategy to utilize the guidance from the sharpest channel and local image regions to improve the overall performance and robustness. Extensive evaluation on both real and synthetic data suggests that the proposed method outperforms the state-of-the-art techniques.
****

Ontological Supervision for Fine Grained Classification of Street View Storefronts
Yair Movshovitz-Attias, Qian Yu, Martin C. Stumpe, Vinay Shet, Sacha Arnoud, Liron Yatziv; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1693-1702
Modern search engines receive large numbers of business related, local aware queries. Such queries are best answered using accurate, up-to-date, business listings, that contain representations of business categories. Creating such listings is a challenging task as businesses often change hands or close down. For businesses with street side locations one can leverage the abundance of street level imagery, such as Google Street View, to automate the process. However, while data is abundant, labeled data is not; the limiting factor is creation of large scale labeled training data. In this work, we utilize an ontology of geographical concepts to automatically propagate business category information and create a large, multi label, training data for fine grained storefront classification. Our learner, which is based on the GoogLeNet/inception Deep Convolutional Network architecture and classifies 208 categories, achieves human level accuracy.
****

Finding Distractors In Images
Ohad Fried, Eli Shechtman, Dan B. Goldman, Adam Finkelstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1703-1712
We propose a new computer vision task we call "distractor prediction." Distractors are the regions of an image that draw attention away from the main subjects and reduce the overall image quality. Removing distractors --- for example, using in-painting --- can improve the composition of an image. In this work we created two datasets of images with user annotations to identify the characteristics of distractors. We use these datasets to train an algorithm to predict distractor maps. Finally, we use our predictor to automatically enhance images.
****

From Image-Level to Pixel-Level Labeling With Convolutional Networks
Pedro O. Pinheiro, Ronan Collobert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1713-1721
We are interested in inferring object segmentation by leveraging only object class information, and by considering only minimal priors on the object segmentation task. This problem could be viewed as a kind of weakly supervised segmentation task, and naturally fits the Multiple Instance Learning (MIL) framework: every training image is known to have (or not) at least one pixel corresponding to the image class label, and the segmentation task can be rewritten as inferring the pixels belonging to the class of the object (given one image, and its object class). We propose a Convolutional Neural Network-based model, which is constrained during training to put more weight on pixels which are important for classifying the image. We show that at test time, the model has learned to discriminate the right pixels well enough, such that it performs very well on an existing segmentation benchmark, by adding only few smoothing priors. Our system is trained using a subset of the Imagenet dataset and the segmentation experiments are performed on the challenging Pascal VOC dataset (with no fine-tuning of the model on Pascal VOC). Our model beats the state of the art results in weakly supervised object segmentation task by a large margin. We also compare the performance of our model with state of the art fully-supervised segmentation approaches.
****

Semantic Alignment of LiDAR Data at City Scale
Fisher Yu, Jianxiong Xiao, Thomas Funkhouser; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1722-1731
This paper describes an automatic algorithm for global alignment of LiDAR data collected with Google Street View cars in urban environments. The problem is challenging because global pose estimation techniques (GPS) do not work well in city environments with tall buildings, and local tracking techniques (integration of inertial sensors, structure-from-motion, etc.) provide solutions that drift over long ranges, leading to solutions where data collected over wide ranges is war

ped and misaligned by many meters. Our approach to address this problem is to e
xtract ``semantic features'' with object detectors (e.g., for facades, poles, ca
rs, etc.) that can be matched robustly at different scales, and thus are selecte
d for different iterations of an ICP algorithm. We have implemented an all-to-a
ll, non-rigid, global alignment based on this idea that provides better results
 than alternatives during experiments with data from large regions of New York, S
an Francisco, Paris, and Rome.
****

Oriented Edge Forests for Boundary Detection
Sam Hallman, Charless C. Fowlkes; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2015, pp. 1732-1740
We present a simple, efficient model for learning boundary detection based on a
 random forest classifier. Our approach combines (1) efficient clustering of trai
ning examples based on a simple partitioning of the space of local edge orientat
ions and (2) scale-dependent calibration of individual tree output probabilities
 prior to multiscale combination. The resulting model outperforms published resu
lts on the challenging BSDS500 boundary detection benchmark. Further, on large d
atasets our model requires substantially less memory for training and speeds up
 training time by a factor of 10 over the structured forest model.
****

Query-Adaptive Late Fusion for Image Search and Person Re-Identification
Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, Qi Tian; Proceedings o
f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, p
p. 1741-1750
Feature fusion has been proven effective [31, 32] in image search. Typically, it
 is assumed that the to-be-fused heterogeneous features work well by themselves
 for the query. However, in a more realistic situation, one does not know in adva
nce whether a feature is effective or not for a given query. As a result, it is
 of great importance to identify feature effectiveness in a query-adaptive manner
. Towards this goal, this paper proposes a simple yet effective late fusion met
hod at score level. Our motivation is that the sorted score curve exhibits an "L
" shape for a good feature, but descends gradually for a bad one (Fig. 1). By ap
proximating score curve's tail with a reference collected on irrelevant data, th
e effectiveness of a feature can be estimated as negatively related to the area
 under the normalized score curve. Experiments are conducted on two image search
 datasets and one person re-identification dataset. We show that our method is r
obust to parameter changes, and outperforms two popular fusion schemes, especial
ly on the resistance to bad features. On the three datasets, our results are com
petitive to the state-of-the-arts.
****

Filtered Feature Channels for Pedestrian Detection
Shanshan Zhang, Rodrigo Benenson, Bernt Schiele; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1751-1760
This paper starts from the observation that multiple top performing pedestrian d
etectors can be modelled by using an intermediate layer filtering low-level feat
ures in combination with a boosted decision forest. Based on this observation we
 propose a unifying framework and experimentally explore different filter famili
es. We report extensive results enabling a systematic analysis. Using filtered
 channel features we obtain top performance on the challenging Caltech and KITTI
 datasets, while using only HOG+LUV as low-level features. When adding optical fl
ow features we further improve detection quality and report the best known resul
t on the Caltech dataset, reaching 93% recall at 1 FPPI.
****

GRSA: Generalized Range Swap Algorithm for the Efficient Optimization of MRFs
Kangwei Liu, Junge Zhang, Peipei Yang, Kaiqi Huang; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1761-1769
Markov Random Field (MRF) is an important tool and has been widely used in many
 vision tasks. Thus, the optimization of MRFs is a problem of fundamental importa
nce. Recently, Veskler and Kumar et. al propose the range move algorithms, which
 are one of the most successful solvers to this problem. However, two problems h

ave limited the applicability of previous range move algorithms: 1) They are lim ited in the types of energies they can handle (i.e. only truncated convex functi ons); 2) These algorithms tend to be very slow compared to other graph-cut based algorithms (e.g. a-expansion and ab-swap). In this paper, we propose a generali zed range swap algorithm (GRSA) for efficient optimization of MRFs. To address t he first problem, we extend the GRSA to arbitrary semimetric energies by restric ting the chosen labels in each move so that the energy is submodular on the chos en subset. Furthermore, to feasibly choose the labels satisfying the submodular condition, we provide a sufficient condition of the submodularity. For the secon d problem, unlike previous range move algorithms which execute the set of all po ssible range moves, we dynamically obtain the iterative moves by solving a set c over problem, which greatly reduces the number of moves during the optimization. Experiments show that the GRSA offers a great speedup over previous range swap algorithms, while it obtains competitive solutions.

****

PatchCut: Data-Driven Object Segmentation via Local Shape Transfer

Jimei Yang, Brian Price, Scott Cohen, Zhe Lin, Ming-Hsuan Yang; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1770-1778

Object segmentation is highly desirable for image understanding and editing. Cur rent interactive tools require a great deal of user effort while automatic metho ds are usually limited to images of special object categories or with high color contrast. In this paper, we propose a data-driven algorithm that uses examples to break through these limits. As similar objects tend to share similar local sh apes, we match query image patches with example images in multiscale to enable l ocal shape transfer. The transferred local shape masks constitute a patch-level segmentation solution space and we thus develop a novel cascade algorithm, Patch Cut, for coarse-to-fine object segmentation. In each stage of the cascade, local shape mask candidates are selected to refine the estimated segmentation of the previous stage iteratively with color models. Experimental results on various da tasets (Weizmann Horse, Fashionista, Object Discovery and PASCAL) demonstrate th e effectiveness and robustness of our algorithm.

****

Illumination and Reflectance Spectra Separation of a Hyperspectral Image Meets L ow-Rank Matrix Factorization

Yinqiang Zheng, Imari Sato, Yoichi Sato; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2015, pp. 1779-1787

This paper addresses the illumination and reflectance spectra separation (IRSS) problem of a hyperspectral image captured under general spectral illumination. T he huge amount of pixels in a hypersepctral image poses tremendous challenges on computational efficiency, yet in turn offers greater color variety that might b e utilized to improve separation accuracy and relax the restrictive subspace ill umination assumption in existing works. We show that this IRSS problem can be mo deled into a low-rank matrix factorization problem, and prove that the separatio n is unique up to an unknown scale under the standard low-dimensionality assumpt ion of reflectance. We also develop a scalable algorithm for this separation tas k that works in the presence of model error and image noise. Experiments on both synthetic data and real images have demonstrated that our separation results ar e sufficiently accurate, and can benefit some important applications, such as sp ectra relighting and illumination swapping.

****

Semantic Part Segmentation Using Compositional Model Combining Shape and Appeara nce

Jianyu Wang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2015, pp. 1788-1797

In this paper, we study the problem of semantic part segmentation for animals. T his is more challenging than standard object detection, object segmentation and pose estimation tasks because semantic parts of animals often have similar appea rance and highly varying shapes. To tackle these challenges, we build a mixture of compositional models to represent the object boundary and the boundaries of s

emantic parts. And we incorporate edge, appearance, and semantic part cues into the compositional model. Given part-level segmentation annotation, we develop a novel algorithm to learn a mixture of compositional models under various poses and viewpoints for certain animal classes. Furthermore, a linear complexity algorithm is offered for efficient inference of the compositional model using dynamic programming. We evaluate our method for horse and cow using a newly annotated dataset on Pascal VOC 2010 which has pixelwise part labels. Experimental results demonstrate the effectiveness of our method.
****

A Discriminative CNN Video Representation for Event Detection
Zhongwen Xu, Yi Yang, Alex G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1798-1807
In this paper, we propose a discriminative video representation for event detection over a large scale video dataset when only limited hardware resources are available. The focus of this paper is to effectively leverage deep Convolutional Neural Networks (CNNs) to advance event detection, where only frame level static descriptors can be extracted by the existing CNN toolkits. This paper makes two contributions to the inference of CNN video representation. First, while average pooling and max pooling have long been the standard approaches to aggregating frame level static features, we show that performance can be significantly improved by taking advantage of an appropriate encoding method. Second, we propose using a set of latent concept descriptors as the frame descriptor, which enriches visual information while keeping it computationally affordable. The integration of the two contributions results in a new state-of-the-art performance in event detection over the largest video datasets. Compared to improved Dense Trajectories, which has been recognized as the best video representation for event detection, our new representation improves the Mean Average Precision (mAP) from 27.6% to 36.8% for the TRECVID MEDTest 14 dataset and from 34.0% to 44.6% for the TRECVID MEDTest 13 dataset.
****

24/7 Place Recognition by View Synthesis
Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1808-1817
We address the problem of large-scale visual place recognition for situations where the scene undergoes a major change in appearance, for example, due to illumination (day/night), change of seasons, aging, or structural modifications over time such as buildings built or destroyed. Such situations represent a major challenge for current large-scale place recognition methods. This work has the following three principal contributions. First, we demonstrate that matching across large changes in the scene appearance becomes much easier when both the query image and the database image depict the scene from approximately the same viewpoint. Second, based on this observation, we develop a new place recognition approach that combines (i) an efficient synthesis of novel views with (ii) a compact indexable image representation. Third, we introduce a new challenging dataset of 1,125 camera-phone query images of Tokyo that contain major changes in illumination (day, sunset, night) as well as structural changes in the scene. We demonstrate that the proposed approach significantly outperforms other large-scale place recognition techniques on this challenging data.
****

Understanding Image Virality
Arturo Deza, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1818-1826
Virality of online content on social networking websites is an important but esoteric phenomenon often studied in fields like marketing, psychology and data mining. In this paper we study viral images from a computer vision perspective. We introduce three new image datasets from Reddit, and define a virality score using Reddit metadata. We train classifiers with state-of-the-art image features to predict virality of individual images, relative virality in pairs of images, and the dominant topic of a viral image. We also compare machine performance to

human performance on these tasks. We find that computers perform poorly with lo w level features, and high level information is critical for predicting viralit y. We encode semantic information through relative attributes. We identify the 5 key visual attributes that correlate with virality. We create an attribute-ba sed characterization of images that can predict relative virality with 68.10% a ccuracy (SVM+Deep Relative Attributes) -- better than humans at 60.12%. Finally, we study how human prediction of image virality varies with different `contexts ' in which the images are viewed, such as the influence of neighbouring images, images recently viewed, as well as the image title or caption. This work is a first step in understanding the complex but important phenomenon of image virali ty. Our datasets and annotations will be made publicly available.
****

Book2Movie: Aligning Video Scenes With Book Chapters
Makarand Tapaswi, Martin Bauml, Rainer Stiefelhagen; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1827-1835
Film adaptations of novels often visually display in a few shots what is describ ed in many pages of the source novel. In this paper we present a new problem: to align book chapters with video scenes. Such an alignment facilitates finding di fferences between the adaptation and the original source, and also acts as a bas is for deriving rich descriptions from the novel for the video clips. We propose an efficient method to compute an alignment between book chapters and video sce nes using matching dialogs and character identities as cues. A major considerati on is to allow the alignment to be non-sequential. Our suggested shortest path b ased approach deals with the non-sequential alignments and can be used to determ ine whether a video scene was part of the original book. We create a new data se t involving two popular novel-to-film adaptations with widely varying properties and compare our method against other text-to-video alignment baselines. Using t he alignment, we present a qualitative analysis of describing the video through rich narratives obtained from the novel.
****

3D Model-Based Continuous Emotion Recognition
Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li, Hongan Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 183 6-1845
We propose a real-time 3D model-based method that continuously recognizes dimens ional emotions from facial expressions in natural communications. In our method, 3D facial models are restored from 2D images, which provide crucial clues for t he enhancement of robustness to overcome large changes including out-of-plane he ad rotations, fast head motions and partial facial occlusions. To accurately rec ognize the emotion, a novel random forest-based algorithm which simultaneously i ntegrates two regressions for 3D facial tracking and continuous emotion estimati on is constructed. Moreover, via the reconstructed 3D facial model, temporal inf ormation and user-independent emotion presentations are also taken into account through our image fusion process. The experimental results show that our algorit hm can achieve state-of-the-art result with higher Pearson's correlation coeffic ient of continuous emotion recognition in real time.
****

Learning to Rank in Person Re-Identification With Metric Ensembles
Sakrapee Paisitkriangkrai, Chunhua Shen, Anton van den Hengel; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1 846-1855
We propose an effective structured learning based approach to the problem of per son re-identification which outperforms the current state-of-the-art on most ben chmark data sets evaluated. Our framework is built on the basis of multiple low-level hand-crafted and high-level visual features. We then formulate two optimiz ation algorithms, which directly optimize evaluation measures commonly used in p erson re-identification, also known as the Cumulative Matching Characteristic (C MC) curve. Our new approach is practical to many real-world surveillance applica tions as the re-identification performance can be concentrated in the range of m ost practical importance. The combination of these factors leads to a person re-

identification system which outperforms most existing algorithms. More importantly, we advance state-of-the-art results on person re-identification by improving the rank-1 recognition rates from 40% to 50% on the iLIDS benchmark, 16% to 18% on the PRID2011 benchmark, 43% to 46% on the VIPeR benchmark, 34% to 53% on the CUHK01 benchmark and 21% to 62% on the CUHK03 benchmark.
****

Making Better Use of Edges via Perceptual Grouping
Yonggang Qi, Yi-Zhe Song, Tao Xiang, Honggang Zhang, Timothy Hospedales, Yi Li, Jun Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1856-1865
We propose a perceptual grouping framework that organizes image edges into meaningful structures and demonstrate its usefulness on various computer vision tasks. Our grouper formulates edge grouping as a graph partition problem, where a learning to rank method is developed to encode probabilities of candidate edge pairs. In particular, RankSVM is employed for the first time to combine multiple Gestalt principles as cue for edge grouping. Afterwards, an edge grouping based object proposal measure is introduced that yields proposals comparable to state-of-the-art alternatives. We further show how human-like sketches can be generated from edge groupings and consequently used to deliver state-of-the-art sketch-based image retrieval performance. Last but not least, we tackle the problem of free hand human sketch segmentation by utilizing the proposed grouper to cluster strokes into semantic object parts.
****

Real-Time Joint Estimation of Camera Orientation and Vanishing Points
Jeong-Kyun Lee, Kuk-Jin Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1866-1874
The widely-used approach for estimating camera orientation is to use points at infinity, i.e., vanishing points (VPs). By enforcing the orthogonal constraint between the VPs, called the Manhattan world constraint, a drift-free camera orientation estimation can be achieved. However, in practical applications this approach suffers from many spurious parallel line segments or does not perform in non-Manhattan world scenes. To overcome these limitations, we propose a novel method that jointly estimates the VPs and camera orientation based on sequential Bayesian filtering. The proposed method does not require the Manhattan world assumption, and can perform a highly accurate estimation of camera orientation in real time. In addition, in order to enhance the robustness of the joint estimation, we propose a feature management technique that removes false positives of line clusters and classifies newly detected lines. We demonstrate the superiority of the proposed method through an extensive evaluation using synthetic and real datasets and comparison with other state-of-the-art methods.
****

Sketch-Based 3D Shape Retrieval Using Convolutional Neural Networks
Fang Wang, Le Kang, Yi Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1875-1883
Retrieving 3D models from 2D human sketches has received considerable attention in the areas of graphics, image retrieval, and computer vision. Almost always in state of the art approaches a large amount of ``best views'' are computed for 3D models, with the hope that the query sketch matches one of these 2D projections of 3D models using predefined features.  We argue that this two stage approach (view selection -- matching) is pragmatic but also problematic because the ``best views'' are subjective and ambiguous, which makes the matching inputs obscure. This imprecise nature of matching further makes it challenging to choose features manually. Instead of relying on the elusive concept of ``best views'' and the hand-crafted features, we propose to define our views using a minimalism approach and learn features for both sketches and views. Specifically, we drastically reduce the number of views to only two predefined directions for the whole dataset. Then, we learn two Siamese Convolutional Neural Networks (CNNs), one for the views and one for the sketches. The loss function is defined on the within-domain as well as the cross-domain similarities. Our experiments on three benchmark datasets demonstrate that our method is significantly better than state of the

art approaches, and outperforms them in all conventional metrics.
****

Salient Object Detection via Bootstrap Learning
Na Tong, Huchuan Lu, Xiang Ruan, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1884-1892
We propose a bootstrap learning algorithm for salient object detection in which both weak and strong models are exploited. First, a weak saliency map is constructed based on image priors to generate training samples for a strong model. Second, a strong classifier based on samples directly from an input image is learned to detect salient pixels. Results from multiscale saliency maps are integrated to further improve the detection performance. Extensive experiments on five benchmark datasets demonstrate that the proposed bootstrap learning algorithm performs favorably against the state-of-the-art saliency detection methods. Furthermore, we show that the proposed bootstrap learning approach can be easily applied to other bottom-up saliency models for significant improvement.
****

Towards Open World Recognition
Abhijit Bendale, Terrance Boult; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1893-1902
With the of advent rich classification models and high computational power visual recognition systems have found many operational applications. Recognition in the real world poses multiple challenges that are not apparent in controlled lab environments. The datasets are dynamic and novel categories must be continuously detected and then added. At prediction time, a trained system has to deal with myriad unseen categories. Operational systems require minimum down time, even to learn. To handle these operational issues, we present the problem of Open World recognition and formally define it. We prove that thresholding sums of monotonically decreasing functions of distances in linearly transformed feature space can balance open space risk and empirical risk. Our theory extends existing algorithms for open world recognition. We present a protocol for evaluation of open world recognition systems. We present the Nearest Non-Outlier (NNO) algorithm which evolves model efficiently, adding object categories incrementally while detecting outliers and managing open space risk. We perform experiments on the ImageNet dataset with 1.2M+ images to validate the effectiveness of our method on large scale visual recognition tasks. NNO consistently yields superior results on open world recognition.
****

Data-Driven 3D Voxel Patterns for Object Category Recognition
Yu Xiang, Wongun Choi, Yuanqing Lin, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1903-1911
Despite the great progress achieved in recognizing objects as 2D bounding boxes in images, it is still very challenging to detect occluded objects and estimate the 3D properties of multiple objects from a single image. In this paper, we propose a novel object representation, 3D Voxel Pattern (3DVP), that jointly encodes the key properties of objects including appearance, 3D shape, viewpoint, occlusion and truncation. We discover 3DVPs in a data-driven way, and train a bank of specialized detectors for a dictionary of 3DVPs. The 3DVP detectors are capable of detecting objects with specific visibility patterns and transferring the meta-data from the 3DVPs to the detected objects, such as 2D segmentation mask, 3D pose as well as occlusion or truncation boundaries. The transferred meta-data allows us to infer the occlusion relationship among objects, which in turn provides improved object recognition results. Experiments are conducted on the KITTI detection benchmark and the outdoor-scene dataset. We improve state-of-the-art results on car detection and pose estimation with notable margins (6% in difficult data of KITTI). We also verify the ability of our method in accurately segmenting objects from the background and localizing them in 3D.
****

3D ShapeNets: A Deep Representation for Volumetric Shapes
Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, Jianxiong Xiao; Proceedings of the IEEE Conference on Computer Vision and Patter

n Recognition (CVPR), 2015, pp. 1912-1920
3D shape is a crucial but heavily underutilized cue in today's computer vision s ystems, mostly due to the lack of a good generic shape representation. With the recent availability of inexpensive 2.5D depth sensors (e.g. Microsoft Kinect), i t is becoming increasingly important to have a powerful 3D shape representation in the loop. Apart from category recognition, recovering full 3D shapes from vie w-based 2.5D depth maps is also a critical part of visual understanding. To this end, we propose to represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid, using a Convolutional Deep Belief Netwo rk. Our model, 3D ShapeNets, learns the distribution of complex 3D shapes across different object categories and arbitrary poses from raw CAD data, and discover s hierarchical compositional part representation automatically. It naturally sup ports joint object recognition and shape completion from 2.5D depth maps, and it enables active object recognition through view planning. To train our 3D deep l earning model, we construct ModelNet - a large-scale 3D CAD model dataset. Exten sive experiments show that our 3D deep representation enables significant perfor mance improvement over the-state-of-the-arts in a variety of tasks.
****

Robust Image Alignment With Multiple Feature Descriptors and Matching-Guided Nei ghborhoods
Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1921-1930
This paper addresses two issues hindering the advances in accurate image alignme nt. First, the performance of descriptor-based approaches to image alignment rel ies on the chosen descriptor, but the optimal descriptor typically varies from i mage to image, or even pixel to pixel. Second, the neighborhood structure for sm oothness enforcement is usually predefined before alignment. However, object bou ndaries are often better discovered during alignment. The proposed approach tack les the two issues by adaptive descriptor selection and dynamic neighborhood con struction. Specifically, we associate each pixel to be aligned with an affine tr ansformation, and integrate the learning of the pixel-specific transformations i nto image alignment. The transformations serve as the common domain for descript or fusion, since the local consensus of each descriptor can be estimated by acce ssing the corresponding affine transformation. It allows us to pick the most pla usible descriptor for aligning each pixel. On the other hand, more object-aware neighborhoods can be produced by referencing the consistency between the learned affine transformations of neighboring pixels. The promising results on popular image alignment benchmarks manifests the effectiveness of our approach.
****

Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Jan us Benchmark A
Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Krist en Allen, Patrick Grother, Alan Mah, Anil K. Jain; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1931-1939
Rapid progress in unconstrained face recognition has resulted in a saturation in recognition accuracy for current benchmark datasets. While important for early progress, a chief limitation in most benchmark datasets is the use of a commodit y face detector to select face imagery. The implication of this strategy is rest ricted variations in face pose and other confounding factors. This paper introdu ces the IARPA Janus Benchmark A (IJB-A), a publicly available media in the wild dataset containing 500 subjects with manually localized face images. Key feature s of the IJB-A dataset are: (i) full pose variation, (ii) joint use for face rec ognition and face detection benchmarking, (iii) a mix of images and videos, (iv) wider geographic variation of subjects, (v) protocols supporting both open-set identification (1:N search) and verification (1:1 comparison), (vi) an optional protocol that allows modeling of gallery subjects, and (vii) ground truth eye a nd nose locations. The dataset has been developed using 1,501,267 million crowd sourced annotations. Baseline accuracies for both face detection and face recogn ition from commercial and open source algorithms demonstrate the challenge offer ed by this new unconstrained benchmark.

****
Depth From Shading, Defocus, and Correspondence Using Light-Field Angular Cohere nce
Michael W. Tao, Pratul P. Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, Ravi Ramamoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition (CVPR), 2015, pp. 1940-1948
Light-field cameras are now used in consumer and industrial applications. Recent papers and products have demonstrated practical depth recovery algorithms from a passive single-shot capture. However, current light field capture devices have narrow baselines and constrained spatial resolution; therefore, the accuracy of depth recovery is limited, requiring heavy regularization and producing planar depths that do not resemble the actual geometry. Using shading information is es sential to improve the shape estimation. We develop an improved technique for lo cal shape estimation from defocus and correspondence cues, and show how shading can be used to further refine the depth. Light-field cameras are able to captur e both spatial and angular data, suitable for refocusing. By locally refocusing each spatial pixel to its respective estimated depth, we produce an all-in-focus image where all viewpoints converge onto a point in the scene. Therefore, the a ngular pixels have angular coherence, which exhibits three properties: photo con sistency, depth consistency, and shading consistency. We propose a new framework that uses angular coherence to optimize depth and shading. The optimization fra mework estimates both general lighting in natural scenes and shading to improve depth regularization. Our method outperforms current state-of-the-art light-fiel d depth estimation algorithms in multiple scenarios, including real images.
****
New Insights Into Laplacian Similarity Search
Xiao-Ming Wu, Zhenguo Li, Shih-Fu Chang; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2015, pp. 1949-1957
Graph-based computer vision applications rely critically on similarity metrics w hich compute the pairwise similarity between any pair of vertices on graphs. Thi s paper investigates the fundamental design of commonly used similarity metrics, and provides new insights to guide their use in practice. In particular, we int roduce a family of similarity metrics in the form of (L+\alpha\Lambda)^{-1}, whe re L is the graph Laplacian, \Lambda is a positive diagonal matrix acting as a r egularizer, and \alpha is a positive balancing factor. Such metrics respect grap h topology when \alpha is small, and reproduce well-known metrics such as hittin g times and the pseudo-inverse of graph Laplacian with different regularizer \La mbda. This paper is the first to analyze the important impact of selecting \ Lambda in retrieving the local cluster from a seed. We find that different \Lamb da can lead to surprisingly complementary behaviors: \Lambda = D (degree matrix) can reliably extract the cluster of a query if it is sparser than surrounding c lusters, while \Lambda = I (identity matrix) is preferred if it is denser than s urrounding clusters. Since in practice there is no reliable way to determine the local density in order to select the right model, we propose a new design of \L ambda that automatically adapts to the local density. Experiments on image retri eval verify our theoretical arguments and confirm the benefit of the proposed me tric. We expect the insights of our theory to provide guidelines for more applic ations in computer vision and other domains.
****
Feature-Independent Context Estimation for Automatic Image Annotation
Amara Tariq, Hassan Foroosh; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2015, pp. 1958-1965
Automatic image annotation is a highly valuable tool for image search, retrieval and archival systems. In the absence of an annotation tool, such systems have t o rely on either users' input or large amount of text on the webpage of the imag e, to acquire its textual description. Users may provide insufficient/noisy tags and all the text on the webpage may not be a description or an explanation of t he accompanying image. Therefore, it is of extreme importance to develop efficie nt tools for automatic annotation of images with correct and sufficient tags. Th e context of the image plays a significant role in this process, along with the

content of the image. A suitable quantification of the context of the image may reduce the semantic gap between visual features and appropriate textual description of the image. In this paper, we present an unsupervised feature-independent quantification of the context of the image through tensor decomposition. We incorporate the estimated context as prior knowledge in the process of automatic image annotation. Evaluation of the predicted annotations provides evidence of the effectiveness of our feature-independent context estimation method.
****

Category-Specific Object Reconstruction From a Single Image
Abhishek Kar, Shubham Tulsiani, Joao Carreira, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1966-1974
Object reconstruction from a single image -- in the wild -- is a problem where we can make progress and get meaningful results today. This is the main message of this paper, which introduces an automated pipeline with pixels as inputs and 3D surfaces of various rigid categories as outputs in images of realistic scenes. At the core of our approach are deformable 3D models that can be learned from 2D annotations available in existing object detection datasets, that can be driven by noisy automatic object segmentations and which we complement with a bottom-up module for recovering high-frequency shape details. We perform a comprehensive quantitative analysis and ablation study of our approach using the recently introduced PASCAL 3D+ dataset and show very encouraging automatic reconstructions on PASCAL VOC.
****

Active Sample Selection and Correction Propagation on a Gradually-Augmented Graph
Hang Su, Zhaozheng Yin, Takeo Kanade, Seungil Huh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1975-1983
When data have a complex manifold structure or the characteristics of data evolve over time, it is unrealistic to expect a graph-based semi-supervised learning method to achieve flawless classification given a small number of initial annotations. To address this issue with minimal human interventions, we propose (i) a sample selection criterion used for \textit{active} query of informative samples by minimizing the expected prediction error, and (ii) an efficient {\it correction propagation} method that propagates human correction on selected samples over a {\it gradually-augmented graph} to unlabeled samples without rebuilding the affinity graph. Experimental results conducted on three real world datasets validate that our active sample selection and correction propagation algorithm quickly reaches high quality classification results with minimal human interventions.
****

Efficient and Accurate Approximations of Nonlinear Convolutional Networks
Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1984-1992
This paper aims to accelerate the test-time computation of deep convolutional neural networks (CNNs). Unlike existing methods that are designed for approximating linear filters or linear responses, our method takes the nonlinear units into account. We minimize the reconstruction error of the nonlinear responses, subject to a low-rank constraint which helps to reduce the complexity of filters. We develop an effective solution to this constrained nonlinear optimization problem. An algorithm is also presented for reducing the accumulated error when multiple layers are approximated. A whole-model speedup ratio of 4x is demonstrated on a large network trained for ImageNet, while the top-5 error rate is only increased by 0.9%. Our accelerated model has a comparably fast speed as the "AlexNet", but is 4.7% more accurate.
****

Ranking and Retrieval of Image Sequences From Multiple Paragraph Queries
Gunhee Kim, Seungwhan Moon, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1993-2001
We propose a method to rank and retrieve image sequences from a natural language

text query, consisting of multiple sentences or paragraphs. One of the method's key applications is to visualize visitors' text-only reviews on TRIPADVISOR or YELP, by automatically retrieving the most illustrative image sequences. While most previous work has dealt with the relations between a natural language sentence and an image or a video, our work extends to the relations between paragraphs and image sequences. Our approach leverages the vast user-generated resource of blog posts and photo streams on the Web. We use blog posts as text-image parallel training data that co-locate informative text with representative images that are carefully selected by users. We exploit large-scale photo streams to augment the image samples for retrieval. We design a latent structural SVM framework to learn the semantic relevance relations between text and image sequences. We present both quantitative and qualitative results on the newly created DISNEYLAND dataset.

****

Casual Stereoscopic Panorama Stitching

Fan Zhang, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2002-2010

This paper presents a method for stitching stereoscopic panoramas from stereo images casually taken using a stereo camera. This method addresses three challenges of stereoscopic image stitching: how to handle parallax, how to stitch the left- and right-view panorama consistently, and how to take care of disparity during stitching. This method addresses these challenges using a three-step approach. First, we employ a state-of-the-art stitching algorithm that handles parallax well to stitch the left views of input stereo images and create the left view of the final stereoscopic panorama. Second, we stitch the input disparity maps to obtain the target disparity map for the stereoscopic panorama by solving a Poisson's equation. This target disparity map is optimized such that there are no vertical disparities and the original perceived depth distribution is preserved. Finally, we warp the right views of the input stereo images and stitch them into the right view of the final stereoscopic panorama according to the target disparity map. The stitching of the right views is formulated as a labeling problem that is constrained by the stitching of the left views to make the left- and right-view panorama consistent to avoid retinal rivalry. Our experiments show that our method can effectively stitch casually taken stereo images and produce high-quality stereoscopic panoramas that deliver a pleasant stereoscopic 3D viewing experience.

****

Superpixel Meshes for Fast Edge-Preserving Surface Reconstruction

Andras Bodis-Szomoru, Hayko Riemenschneider, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2011-2020

Multi-View-Stereo (MVS) methods aim for the highest detail possible, however, such detail is often not required. In this work, we propose a novel surface reconstruction method based on image edges, superpixels and second-order smoothness constraints, producing meshes comparable to classic MVS surfaces in quality but orders of magnitudes faster. Our method performs per-view dense depth optimization directly over sparse 3D Ground Control Points (GCPs), hence, removing the need for view pairing, image rectification, and stereo depth estimation, and allowing for full per-image parallelization. We use Structure-from-Motion (SfM) points as GCPs, but the method is not specific to these, e.g.~LiDAR or RGB-D can also be used. The resulting meshes are compact and inherently edge-aligned with image gradients, enabling good-quality lightweight per-face flat renderings. Our experiments demonstrate on a variety of 3D datasets the superiority in speed and competitive surface quality.

****

Best-Buddies Similarity for Robust Template Matching

Tali Dekel, Shaul Oron, Michael Rubinstein, Shai Avidan, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2021-2029

We propose a novel method for template matching in unconstrained environments. I

ts essence is the Best Buddies Similarity (BBS), a useful, robust, and parameter-free similarity measure between two sets of points. BBS is based on a count of Best Buddies Pairs (BBPs)--pairs of points in which each one is the nearest neighbor of the other. BBS has several key features that make it robust against complex geometric deformations and high levels of outliers, such as those arising from background clutter and occlusions. We study these properties, provide a statistical analysis that justifies them, and demonstrate the consistent success of BBS on a challenging real-world dataset.

****

Superdifferential Cuts for Binary Energies

Tatsunori Taniai, Yasuyuki Matsushita, Takeshi Naemura; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2030-2038

We propose an efficient and general purpose energy optimization method for binary variable energies used in various low-level vision tasks. The proposed method can be used for broad classes of higher-order and pairwise non-submodular functions. We first revisit a submodular-supermodular procedure (SSP) [Narasimhan05], which is previously studied for higher-order energy optimization. We then present our method as generalization of SSP, which is further shown to generalize several state-of-the-art techniques for higher-order and pairwise non-submodular functions [Ayed13, Gorelick14, Tang14]. In the experiments, we apply our method to image segmentation, deconvolution, and binarization, and show improvements over state-of-the-art methods.

****

The S-Hock Dataset: Analyzing Crowds at the Stadium

Davide Conigliaro, Paolo Rota, Francesco Setti, Chiara Bassetti, Nicola Conci, Nicu Sebe, Marco Cristani; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2039-2047

The topic of crowd modeling in computer vision usually assumes a single generic typology of crowd, which is very simplistic. In this paper we adopt a taxonomy that is widely accepted in sociology, focusing on a particular category, the spectator crowd, which is formed by people "interested in watching something specific that they came to see". This can be found at the stadiums, amphitheaters, cinema, etc. In particular, we propose a novel dataset, the Spectators Hockey (S-Hock), which deals with 4 hockey matches during an international tournament. In the dataset, a massive annotation has been carried out, focusing on the spectators at different levels of details: at a higher level, people have been labeled depending on the team they are supporting and the fact that they know the people close to them; going to the lower levels, standard pose information has been considered (regarding the head, the body) but also fine grained actions such as hands on hips, clapping hands etc. The labeling focused on the game field also, permitting to relate what is going on in the match with the crowd behavior. This brought to more than 100 millions of annotations, useful for standard applications as people counting and head pose estimation but also for novel tasks as spectator categorization. For all of these we provide protocols and baseline results, encouraging further research.

****

Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition With Image Sets

Wen Wang, Ruiping Wang, Zhiwu Huang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2048-2057

This paper presents a method named Discriminant Analysis on Riemannian manifold of Gaussian distributions (DARG) to solve the problem of face recognition with image sets. Our goal is to capture the underlying data distribution in each set and thus facilitate more robust classification. To this end, we represent image set as Gaussian Mixture Model (GMM) comprising a number of Gaussian components with prior probabilities and seek to discriminate Gaussian components from different classes. In the light of information geometry, the Gaussians lie on a specific Riemannian manifold. To encode such Riemannian geometry properly, we investiga

te several distances between Gaussians and further derive a series of provably p
ositive definite probabilistic kernels. Through these kernels, a weighted Kernel
 Discriminant Analysis is finally devised which treats the Gaussians in GMMs as
samples and their prior probabilities as sample weights. The proposed method is
evaluated by face identification and verification tasks on four most challenging
 and largest databases, YouTube Celebrities, COX, YouTube Face DB and Point-and-
Shoot Challenge, to demonstrate its superiority over the state-of-the-art.
****

Texture Representations for Image and Video Synthesis
Georgios Georgiadis, Alessandro Chiuso, Stefano Soatto; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2058-206
6
In texture synthesis and classification, algorithms require a small texture to b
e provided as an input, which is assumed to be representative of a larger region
 to be re-synthesized or categorized. We focus on how to characterize such textu
res and automatically retrieve them. Most works generate these small input textu
res manually by cropping, which does not ensure maximal compression, nor that th
e selection is the best representative of the original. We construct a new repre
sentation that compactly summarizes a texture, while using less storage, that ca
n be used for texture compression and synthesis. We also demonstrate how the rep
resentation can be integrated in our proposed video texture synthesis algorithm
to generate novel instances of textures and video hole-filling. Finally, we prop
ose a novel criterion that measures structural and statistical dissimilarity bet
ween textures.
****

Shadow Optimization From Structured Deep Edge Detection
Li Shen, Teck Wee Chua, Karianto Leman; Proceedings of the IEEE Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2015, pp. 2067-2074
We present a novel learning-based framework for shadow detection from a single i
mage. The local structure of shadow boundaries as well as the global interaction
s of the shadow and non-shadow regions remain largely unexploited by previous le
arning-based approaches. In this paper, we propose an efficient structured label
ling framework for shadow detection from a single image. A convolutional Neural
Networks framework is designed to capture the local structure information of sha
dow edge and to learn the most relevant features. We further propose and formula
te a global shadow optimization framework which can model the complex global int
eractions over the shadow and light regions. Using the shadow edges detected by
our proposed method, the shadow map can be solved by efficient least-square opti
mization. Our proposed framework is efficient and achieves state-of-the-art resu
lts on the major shadow benchmark databases collected under a variety of conditi
ons.
****

Total Variation Regularization of Shape Signals
Maximilian Baust, Laurent Demaret, Martin Storath, Nassir Navab, Andreas Weinman
n; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
 (CVPR), 2015, pp. 2075-2083
This paper introduces the concept of shape signals, i.e., series of shapes which
 have a natural temporal or spatial ordering, as well as a variational formulati
on for the regularization of these signals. The proposed formulation can be seen
 as the shape-valued generalization of the Rudin-Osher-Fatemi (ROF) functional f
or intensity images. We derive a variant of the classical finite-dimensional rep
resentation of Kendall, but our framework is generic in the sense that it can be
 combined with any shape space. This representation allows for the explicit comp
utation of geodesics and thus facilitates the efficient numerical treatment of t
he variational formulation by means of the cyclic proximal point algorithm. Simi
lar to the ROF-functional, we demonstrate experimentally that l_1-type penalties
 both for data fidelity term and regularizer perform best in regularizing shape
signals. Finally, we show applications of our method to shape signals obtained f
rom synthetic, photometric, and medical data sets.
****

Learning Similarity Metrics for Dynamic Scene Segmentation
Damien Teney, Matthew Brown, Dmitry Kit, Peter Hall; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2084-2093
This paper addresses the segmentation of videos with arbitrary motion, including dynamic textures, using novel motion features and a supervised learning approach. Dynamic textures are commonplace in natural scenes, and exhibit complex patterns of appearance and motion (e.g. water, smoke, swaying foliage). These are difficult for existing segmentation algorithms, often violate the brightness constancy assumption needed for optical flow, and have complex segment characteristics beyond uniform appearance or motion. Our solution uses custom spatiotemporal filters that capture texture and motion cues, along with a novel metric-learning framework that optimizes this representation for specific objects and scenes. This is used within a hierarchical, graph-based segmentation setting, yielding state-of-the-art results for dynamic texture segmentation. We also demonstrate the applicability of our approach to general object and motion segmentation, showing significant improvements over unsupervised segmentation and results comparable to the best task specific approaches.
****

Subspace Clustering by Mixture of Gaussian Regression
Baohua Li, Ying Zhang, Zhouchen Lin, Huchuan Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2094-2102
Subspace clustering is a problem of finding a multisubspace representation that best fits sample points drawn from a high-dimensional space. The existing clustering models generally adopt different norms to describe noise, which is equivalent to assuming that the data are corrupted by specific types of noise. In practice, however, noise is much more complex. So it is inappropriate to simply use a certain norm to model noise. Therefore, we propose Mixture of Gaussian Regression (MoG Regression) for subspace clustering by modeling noise as a Mixture of Gaussians (MoG). The MoG Regression provides an effective way to model a much broader range of noise distributions. As a result, the obtained affinity matrix is better at characterizing the structure of data in real applications. Experimental results on multiple datasets demonstrate that MoG Regression significantly outperforms state-of-the-art subspace clustering methods.
****

DASC: Dense Adaptive Self-Correlation Descriptor for Multi-Modal and Multi-Spectral Correspondence
Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N. Do, Kwanghoon Sohn; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2103-2112
Establishing dense visual correspondence between multiple images is a fundamental task in many applications of computer vision and computational photography. Classical approaches, which aim to estimate dense stereo and optical flow fields for images adjacent in viewpoint or in time, have been dramatically advanced in recent studies. However, finding reliable visual correspondence in multi-modal or multi-spectral images still remains unsolved. In this paper, we propose a new dense matching descriptor, called dense adaptive self-correlation (DASC), to effectively address this kind of matching scenarios. Based on the observation that a self-similarity existing within images is less sensitive to modality variations, we define a novel descriptor with a series of an adaptive self-correlation similarity for patches within a local support window. To further improve the matching quality and runtime efficiency, we propose a patch-wise receptive field pooling, in which a sampling pattern is optimized with a discriminative learning. Moreover, the computational redundancy that arises when computing densely sampled descriptor over an entire image is dramatically reduced by applying fast edge-aware filtering. Experiments demonstrate the outstanding performance of the DASC descriptor in many cases of multi-modal and multi-spectral correspondence.
****

In Defense of Color-Based Model-Free Tracking
Horst Possegger, Thomas Mauthner, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2113-2120

In this paper, we address the problem of model-free online object tracking based on color representations. According to the findings of recent benchmark evaluations, such trackers often tend to drift towards regions which exhibit a similar appearance compared to the object of interest. To overcome this limitation, we propose an efficient discriminative object model which allows us to identify potentially distracting regions in advance. Furthermore, we exploit this knowledge to adapt the object representation beforehand so that distractors are suppressed and the risk of drifting is significantly reduced. We evaluate our approach on recent online tracking benchmark datasets demonstrating state-of-the-art results. In particular, our approach performs favorably both in terms of accuracy and robustness compared to recent tracking algorithms. Moreover, the proposed approach allows for an efficient implementation to enable online object tracking in real-time.
****

Best of Both Worlds: Human-Machine Collaboration for Object Annotation
Olga Russakovsky, Li-Jia Li, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2121-2131
The long-standing goal of localizing every object in an image remains elusive. Manually annotating objects is quite expensive despite crowd engineering innovations. Current state-of-the-art automatic object detectors can accurately detect at most a few objects per image. This paper brings together the latest advancements in object detection and in crowd engineering into a principled framework for accurately and efficiently localizing objects in images. The input to the system is an image to annotate and a set of annotation constraints: desired precision, utility and/or human cost of the labeling. The output is a set of object annotations, informed by human feedback and computer vision. Our model seamlessly integrates multiple computer vision models with multiple sources of human input in a Markov Decision Process. We empirically validate the effectiveness of our human-in-the-loop labeling approach on the ILSVRC2014 object detection dataset.
****

Robust Multiple Homography Estimation: An Ill-Solved Problem
Zygmunt L. Szpak, Wojciech Chojnacki, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2132-2141
The estimation of multiple homographies between two piecewise planar views of a rigid scene is often assumed to be a solved problem. We show that contrary to popular opinion various crucial aspects of the task have not been adequately emphasised. We are motivated by a growing body of literature in robust multi-structure estimation that purports to solve the multi-homography estimation problem but in fact does not. We demonstrate that the estimation of multiple homographies is an ill-solved problem by deriving new constraints that a set of mutually compatible homographies must satisfy, and by showing that homographies estimated with prevailing methods fail to satisfy the requisite constraints on real-world data. We also explain why incompatible homographies imply inconsistent epipolar geometries. The arguments and experiments presented in this paper signal the need for a new generation of robust multi-structure estimation methods that have the capacity to enforce constraints on projective entities such as homography matrices.
****

Semi-Supervised Domain Adaptation With Subspace Learning for Visual Recognition
Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, Tao Mei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2142-2150
In many real-world applications, we are often facing the problem of cross domain learning, i.e., to borrow the labeled data or transfer the already learnt knowledge from a source domain to a target domain. However, simply applying existing source data or knowledge may even hurt the performance, especially when the data distribution in the source and target domain is quite different, or there are very few labeled data available in the target domain. This paper proposes a novel domain adaptation framework, named Semi-supervised Domain Adaptation with Subspace Learning (SDASL), which jointly explores invariant low-dimensional structure

s across domains to correct data distribution mismatch and leverages available u
nlabeled target examples to exploit the underlying intrinsic information in the
target domain. Specifically, SDASL conducts the learning by simultaneously minim
izing the classification error, preserving the structure within and across domai
ns, and restricting similarity defined on unlabeled target examples. Encouraging
 results are reported for two challenging domain transfer tasks (including image
-to-image and image-to-video transfers) on several standard datasets in the cont
ext of both image object recognition and video concept detection.
****

Articulated Motion Discovery Using Pairs of Trajectories
Luca Del Pero, Susanna Ricco, Rahul Sukthankar, Vittorio Ferrari; Proceedings of
 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp
. 2151-2160
We propose an unsupervised approach for discovering characteristic motion patter
ns in videos of highly articulated objects performing natural, unscripted behavi
ors, such as tigers in the wild. We discover consistent patterns in a bottom-up
manner by analyzing the relative displacements of large numbers of ordered traje
ctory pairs through time, such that each trajectory is attached to a different m
oving part on the object. The pairs of trajectories descriptor relies entirely o
n motion and is more discriminative than state-of-the-art features that employ s
ingle trajectories. Our method generates temporal video intervals, each automati
cally trimmed to one instance of the discovered behavior, and clusters them by t
ype (e.g., running, turning head, drinking water). We present experiments on two
 datasets: dogs from YouTube-Objects and a new dataset of National Geographic ti
ger videos. Results confirm that our proposed descriptor outperforms existing ap
pearance- and trajectory-based descriptors (e.g., HOG and DTFs) on both datasets
 and enables us to segment unconstrained animal video into intervals containing
single behaviors.
****

A Solution for Multi-Alignment by Transformation Synchronisation
Florian Bernard, Johan Thunberg, Peter Gemmar, Frank Hertel, Andreas Husch, Jorg
e Goncalves; Proceedings of the IEEE Conference on Computer Vision and Pattern R
ecognition (CVPR), 2015, pp. 2161-2169
The alignment of a set of objects by means of transformations plays an important
 role in computer vision. Whilst the case for only two objects can be solved glo
bally, when multiple objects are considered usually iterative methods are used.
In practice the iterative methods perform well if the relative transformations b
etween any pair of objects are free of noise. However, if only noisy relative tr
ansformations are available (e.g. due to missing data or wrong correspondences)
the iterative methods may fail. Based on the observation that the underlying no
ise-free transformations can be retrieved from the null space of a matrix that c
an directly be obtained from pairwise alignments, this paper presents a novel me
thod for the synchronisation of pairwise transformations such that they are tran
sitively consistent. Simulations demonstrate that for noisy transformations, a
 large proportion of missing data and even for wrong correspondence assignments
the method delivers encouraging results.
****

A Convex Optimization Approach to Robust Fundamental Matrix Estimation
Yongfang Cheng, Jose A. Lopez, Octavia Camps, Mario Sznaier; Proceedings of the
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 217
0-2178
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 constrained polynomial optimization, for which a
  monotonically convergent sequence of tractable convex relaxations can be obtai
ned by exploiting recent developments in sparse polynomial optimization. Furthe
r, these results allow for deriving conditions certifying that a finite order re
laxation has converged to a solution. A salient feature of the proposed approac
h is its ability to incorporate existing a-priori information about the noise,
co-ocurrences, and percentage of outliers. These results are illustrated with s

everal examples where the proposed algorithm is shown to outperform existing app
roaches.
****

Simultaneous Pose and Non-Rigid Shape With Particle Dynamics
Antonio Agudo, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Com
puter Vision and Pattern Recognition (CVPR), 2015, pp. 2179-2187
In this paper, we propose a sequential solution to simultaneously estimate camer
a pose and non-rigid 3D shape from a monocular video. In contrast to most existi
ng approaches that rely on global representations of the shape, we model the obj
ect at a local level, as an ensemble of particles, each ruled by the linear equa
tion of the Newton's second law of motion. This dynamic model is incorporated in
to a bundle adjustment framework, in combination with simple regularization comp
onents that ensure temporal and spatial consistency of the estimated shape and c
amera poses. The resulting approach is both efficient and robust to several arti
facts such as noisy and missing data or sudden camera motions, while it does not
 require any training data at all. Validation is done in a variety of real video
 sequences, including articulated and non-rigid motion, both for continuous and
discontinuous shapes. Our system is shown to perform comparable to competing bat
ch, computationally expensive, methods and shows remarkable improvement with res
pect to the sequential ones.
****

Semi-Supervised Learning With Explicit Relationship Regularization
Kwang In Kim, James Tompkin, Hanspeter Pfister, Christian Theobalt; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015,
pp. 2188-2196
In many learning tasks, the structure of the target space of a function holds ri
ch information about the relationships between evaluations of functions on diffe
rent data points. Existing approaches attempt to exploit this relationship infor
mation implicitly by enforcing smoothness on function evaluations only. However,
 what happens if we explicitly regularize the relationships between function eva
luations? Inspired by homophily, we regularize based on a smooth relationship fu
nction, either defined from the data or with labels. In experiments, we demonstr
ate that this significantly improves the performance of state-of-the-art algorit
hms in semi-supervised classification and in spectral data embedding for constra
ined clustering and dimensionality reduction.
****

Person Re-Identification by Local Maximal Occurrence Representation and Metric L
earning
Shengcai Liao, Yang Hu, Xiangyu Zhu, Stan Z. Li; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2197-2206
Person re-identification is an important technique towards automatic search of a
 person's presence in a surveillance video. Two fundamental problems are critica
l for person re-identification, feature representation and metric learning. An e
ffective feature representation should be robust to illumination and viewpoint c
hanges, and a discriminant metric should be learned to match various person imag
es. In this paper, we propose an effective feature representation called Local M
aximal Occurrence (LOMO), and a subspace and metric learning method called Cross
-view Quadratic Discriminant Analysis (XQDA). The LOMO feature analyzes the hori
zontal occurrence of local features, and maximizes the occurrence to make a stab
le representation against viewpoint changes. Besides, to handle illumination var
iations, we apply the Retinex transform and a scale invariant texture operator.
To learn a discriminant metric, we propose to learn a discriminant low dimension
al subspace by cross-view quadratic discriminant analysis, and simultaneously, a
 QDA metric is learned on the derived subspace. We also present a practical comp
utation method for XQDA, as well as its regularization. Experiments on four chal
lenging person re-identification databases, VIPeR, QMUL GRID, CUHK Campus, and C
UHK03, show that the proposed method improves the state-of-the-art rank-1 identi
fication rates by 2.2%, 4.88%, 28.91%, and 31.55% on the four databases, respect
ively.
****

Joint Patch and Multi-Label Learning for Facial Action Unit Detection
Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, Honggang Zhang ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2207-2216
The face is one of the most powerful channel of non-verbal communication. The most commonly used taxonomy to describe facial behaviour is the Facial Action Coding System (FACS). FACS segments the visible effects of facial muscle activation into 30+ action units (AUs). AUs, which may occur alone and in thousands of combinations, can describe nearly all-possible facial expressions. Most existing methods for automatic AU detection treat the problem using one-vs-all classifiers and fail to exploit dependencies among AU and facial features. We introduce joint-patch and multi-label learning (JPML) to address these issues. JPML leverages group sparsity by selecting a sparse subset of facial patches while learning a multi-label classifier. In four of five comparisons on three diverse datasets, CK+, GFT, and BP4D, JPML produced the highest average F1 scores in comparison with state-of-the art.
****

Real-Time Visual Analysis of Microvascular Blood Flow for Critical Care
Chao Liu, Hernando Gomez, Srinivasa Narasimhan, Artur Dubrawski, Michael R. Pinsky, Brian Zuckerbraun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2217-2225
Microcirculatory monitoring plays an important role in diagnosis and treatment of critical care patients. Sidestream Dark Field (SDF) imaging devices have been used to visualize and support interpretation of the micro-vascular blood flow. However, due to subsurface scattering within the tissue that embeds the capillaries, transparency of plasma, imaging noise and lack of features, it is difficult to obtain reliable physiological data from SDF videos. Therefore, thus far microcirculatory videos have been analyzed manually with significant input from expert clinicians. In this paper, we present a framework that automates the analysis process. It includes stages of video stabilization, enhancement, and micro-vessel extraction, in order to automatically estimate statistics of the micro blood flows from SDF videos. Our method has been validated in critical care experiments conducted carefully to record the microcirculatory blood flow in test animal subjects before, during and after induced bleeding episodes, as well as to study the effect of fluid resuscitation. Our method is able to extract microcirculatory measurements that are consistent with clinical intuition and it has a potential to become a useful tool in critical care medicine.
****

JOTS: Joint Online Tracking and Segmentation
Longyin Wen, Dawei Du, Zhen Lei, Stan Z. Li, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2226-2234
We present a novel Joint Online Tracking and Segmentation (JOTS) algorithm which integrates the multi-part tracking and segmentation into a unified energy optimization framework to handle the video segmentation task. The multi-part segmentation is posed as a pixel-level label assignment task with regularization according to the estimated part models, and tracking is formulated as estimating the part models based on the pixel labels, which in turn is used to refine the model. The multi-part tracking and segmentation are carried out iteratively to minimize the proposed objective function by a RANSAC-style approach. Extensive experiments on the SegTrack and SegTrack v2 databases demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.
****

Gaze-Enabled Egocentric Video Summarization via Constrained Submodular Maximization
Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, Vikas Singh ; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2235-2244
With the proliferation of wearable cameras, the number of videos of users documenting their personal lives using such devices is rapidly increasing. Since suc

h videos may span hours, there is an important need for mechanisms that represent the information content in a compact form (i.e., shorter videos which are more easily browsable/sharable). Motivated by these applications, this paper focuses on the problem of egocentric video summarization. Such videos are usually continuous with significant camera shake and other quality issues. Because of these reasons, there is growing consensus that direct application of standard video summarization tools to such data yields unsatisfactory performance. In this paper, we demonstrate that using gaze tracking information (such as fixation and saccade) significantly helps the summarization task. It allows meaningful comparison of different image frames and enables deriving personalized summaries (gaze provides a sense of the camera wearer's intent). We formulate a summarization model which captures common-sense properties of a good summary, and show that it can be solved as a submodular function maximization with partition matroid constraints, opening the door to a rich body of work from combinatorial optimization. We evaluate our approach on a new gaze-enabled egocentric video dataset (over 15 hours), which will be a valuable standalone resource.
****

Sparse Depth Super Resolution
Jiajun Lu, David Forsyth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2245-2253
We describe a method to produce detailed high resolution depth maps from aggressively subsampled depth measurements. Our method fully uses the relationship between image segmentation boundaries and depth boundaries. It uses an image combined with a low resolution depth map. 1) The image is segmented with the guidance of sparse depth samples. 2) Each segment has its depth field reconstructed independently using a novel smoothing method. 3) For videos, time-stamped samples from near frames are incorporated. The paper shows reconstruction results of super resolution from x4 to x100, while previous methods mainly work on x2 to x16. The method is tested on four different datasets and six video sequences, covering quite different regimes, and it outperforms recent state of the art methods quantitatively and qualitatively. We also demonstrate that depth maps produced by our method can be used by applications such as hand trackers, while depth maps from other methods have problems.
****

Efficient Illuminant Estimation for Color Constancy Using Grey Pixels
Kai-Fu Yang, Shao-Bing Gao, Yong-Jie Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2254-2263
Illuminant estimation is a key step for computational color constancy. Instead of using the grey world or grey edge assumptions, we propose in this paper a novel method for illuminant estimation by using the information of grey pixels detected in a given color-biased image. The underlying hypothesis is that most of the natural images include some detectable pixels that are at least approximately grey, which can be reliably utilized for illuminant estimation. We first validate our assumption through comprehensive statistical evaluation on diverse collection of datasets and then put forward a novel grey pixel detection method based on the illuminant-invariant measure (IIM) in three logarithmic color channels. Then the light source color of a scene can be easily estimated from the detected grey pixels. Experimental results on four benchmark datasets (three recorded under single illuminant and one under multiple illuminants) show that the proposed method outperforms most of the state-of-the-art color constancy approaches with the inherent merit of low computational cost.
****

Can Humans Fly? Action Understanding With Multiple Classes of Actors
Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, Jason J. Corso; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2264-2273
Can humans fly? Emphatically no. Can cars eat? Again, absolutely not. Yet, these absurd inferences result from the current disregard for particular types of actors in action understanding. There is no work we know of on simultaneously inferring actors and actions in the video, not to mention a dataset to experiment wit

h. Our paper hence marks the first effort in the computer vision community to jo intly consider various types of actors undergoing various actions. To start with the problem, we collect a dataset of 3782 videos from YouTube and label both pi xel-level actors and actions in each video. We formulate the general actor-actio n understanding problem and instantiate it at various granularities: both video- level single- and multiple-label actor-action recognition and pixel-level actor- action semantic segmentation. Our experiments demonstrate that inference jointly over actors and actions outperforms inference independently over them, and henc e concludes our argument of the value of explicit consideration of various actor s in comprehensive action understanding.
****

Reweighted Laplace Prior Based Hyperspectral Compressive Sensing for Unknown Spa rsity
Lei Zhang, Wei Wei, Yanning Zhang, Chunna Tian, Fei Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2274-228 1
Compressive sensing(CS) has been exploited for hypespectral image(HSI) compressi on in recent years. Though it can greatly reduce the costs of computation and st orage, the reconstruction of HSI from a few linear measurements is challenging. The underlying sparsity of HSI is crucial to improve the reconstruction accuracy . However, the sparsity of HSI is unknown in reality and varied with different n oise, which makes the sparsity estimation difficult. To address this problem, a novel reweighted Laplace prior based hyperspectral compressive sensing method is proposed in this study. First, the reweighted Laplace prior is proposed to mode l the distribution of sparsity in HSI. Second, the latent variable Bayes model i s employed to learn the optimal configuration of the reweighted Laplace prior fr om the measurements. The model unifies signal recovery, prior learning and noise estimation into a variational framework to infer the parameters automatically. The learned sparsity prior can represent the underlying structure of the sparse signal very well and is adaptive to the unknown noise, which improves the recons truction accuracy of HSI. The experimental results on three hyperspectral datase ts demonstrate the proposed method outperforms serveral state-of-the-art hypersp ectral CS methods on the reconstruction accuracy.
****

Class Consistent Multi-Modal Fusion With Binary Features
Ashish Shrivastava, Mohammad Rastegari, Sumit Shekhar, Rama Chellappa, Larry S. Davis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni tion (CVPR), 2015, pp. 2282-2291
Many existing recognition algorithms combine different modalities based on train ing accuracy but do not consider the possibility of noise at test time. We descr ibe an algorithm that perturbs test features so that all modalities predict the same class. We enforce this perturbation to be as small as possible via a quadra tic program (QP) for continuous features, and a mixed integer program (MIP) for binary features. To efficiently solve the MIP, we provide a greedy algorithm and empirically show that its solution is very close to that of a state-of-the-art MIP solver. We evaluate our algorithm on several datasets and show that the meth od outperforms existing approaches.
****

R6P - Rolling Shutter Absolute Camera Pose
Cenek Albl, Zuzana Kukelova, Tomas Pajdla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2292-2300
We present a minimal, non-iterative solution to the absolute pose problem for im ages from rolling shutter cameras. Absolute pose problem is a key problem in com puter vision and rolling shutter is present in a vast majority of today's digita l cameras. We propose several rolling shutter camera models and verify their fea sibility for a polynomial solver. A solution based on linearized camera model is chosen and verified in several experiments. We use a linear approximation to th e camera orientation, which is meaningful only around the identity rotation. We show that the standard P3P algorithm is able to estimate camera orientation with in 6 degrees for camera rotation velocity as high as 30deg/frame. Therefore we c

an use the standard P3P algorithm to estimate camera orientation and to bring th e camera rotation matrix close to the identity. Using this solution, camera posi tion, orientation, translational velocity and angular velocity can be computed u sing six 2D-to-3D correspondences, with orientation error under half a degree an d relative position error under 2%. A significant improvement in terms of the nu mber of inliers in RANSAC is demonstrated.
****

Embedded Phase Shifting: Robust Phase Shifting With Embedded Signals
Daniel Moreno, Kilho Son, Gabriel Taubin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2301-2309
We introduce Embedded PS, a new robust and accurate phase shifting algorithm for 3D scanning. The method projects only high frequency sinusoidal patterns in ord er to reduce errors due to global illumination effects, such as subsurface scatt ering and interreflections. The frequency set for the projected patterns is spec ially designed so that our algorithm can extract a set of embedded low frequency sinusoidals with simple math. All the signals, patterns high and embedded low f requencies, are used with temporal phase unwrapping to compute absolute phase va lues in closed-form, without quantization or approximation via LUT, resulting in fast computation. The absolute phases provide correspondences from projector to camera pixels which enable to recover 3D points using optical triangulation. Th e algorithm estimates multiple absolute phase values per pixel which are combine d to reduce measurement noise while preserving fine details. We prove that embed ded periodic signals can be recovered from any periodic signal, not just sinusoi dal signals, which may result in further improvements for other 3D imaging metho ds. Several experiments are presented showing that our algorithm produces more r obust and accurate 3D scanning results than state-of-the-art methods for challen ging surface materials, with an equal or smaller number of projected patterns an d at lower computational cost.
****

Shape and Light Directions From Shading and Polarization
Trung Ngo Thanh, Hajime Nagahara, Rin-ichiro Taniguchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2310-231 8
We introduce a method to recover the shape of a smooth dielectric object from po larization images taken with a light source from different directions. We presen t two constraints on shading and polarization and use both in a single optimizat ion scheme. This integration is motivated by the fact that photometric stereo an d polarization-based methods have complementary abilities. The polarization-base d method can give strong cues for the surface orientation and refractive index, which are independent of the light direction. However, it has ambiguities in sel ecting between two ambiguous choices of the surface orientation, in the relation ship between refractive index and zenith angle (observing angle), and limited pe rformance for surface points with small zenith angles, where the polarization ef fect is weak. In contrast, photometric stereo method with multiple light sources can disambiguate the surface orientation and give a strong relationship between the surface normals and light directions. However, it has limited performance f or large zenith angles, refractive index estimation, and faces the ambiguity in case the light direction is unknown. Taking their advantages, our proposed metho d can recover the surface normals for both small and large zenith angles, the li ght directions, and the refractive indexes of the object. The proposed method is successfully evaluated by simulation and real-world experiments.
****

3D Deep Shape Descriptor
Yi Fang, Jin Xie, Guoxian Dai, Meng Wang, Fan Zhu, Tiantian Xu, Edward Wong; Pro ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2319-2328
Shape descriptor is a concise yet informative representation that provides a 3D object with an identification as a member of some category. This paper developed a concise deep shape descriptor for the first time to address challenging issue s from ever-growing 3D datasets in areas as diverse as engineering, medicine, an

d biology. Specifically, the proposed approach developed novel techniques to ext
ract concise but geometrically informative shape descriptor, new definitions of
Eigen-shape descriptor and Fisher-shape descriptor to guide the training strateg
y for deep neural network, and deep shape descriptor with discriminative capacit
y of maximizing the inter-class margin while minimizing the intra-class variance
. Our approach addressed the challenges for shape analysis techniques posed by t
he complexity of 3D model and data representation and geometric structural varia
tions and noise present in 3D models. The experimental results on 3D shape retri
eval demonstrate that our proposed deep shape descriptor is superior to other st
ate-of-the-art approaches on handling noise, incompleteness and 3D shape structu
ral variations.
****

Cross-Age Face Verification by Coordinating With Cross-Face Age Verification
Liang Du, Haibin Ling; Proceedings of the IEEE Conference on Computer Vision and
 Pattern Recognition (CVPR), 2015, pp. 2329-2338
In this paper we present a novel framework for cross-age face verification (FV)
by seeking help from its ``competitor" named cross-face age verification (AV), i
.e., deciding whether two face photos are taken at similar ages. While FV and AV
 share some common features, FV pursues age insensitivity and AV seeks age sensi
tivity. Such correlation suggests that AV may be used to guide feature selectio
n in FV, i.e., by reducing the chance of choosing age sensitive features. Dri
ven by this intuition, we propose to learn a solution for cross-age face verific
ation by coordinating with a solution for age verification. Specifically, a join
t additive model is devised to simultaneously handling both tasks, while encodin
g feature coordination by a competition regularization term. Then, an alterna
ting greedy coordinate descent (AGCD) algorithm is developed to solve this joint
 model. As shown in our experiments, the algorithm effectively balances feature
 sharing and feature exclusion between the two tasks; and, for face verification
, the algorithm effectively removes distracting features used in age verificatio
n. To evaluate the proposed algorithm, we conduct cross-age face verification
 experiments using two benchmark cross-age face datasets, FG-Net and MORPH. In a
ll experiments, our algorithm achieves very promising results and outperforms al
l previously tested solutions.
****

Beyond Mahalanobis Metric: Cayley-Klein Metric Learning
Yanhong Bi, Bin Fan, Fuchao Wu; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2015, pp. 2339-2347
Cayley-Klein metric is a kind of non-Euclidean metric suitable for projective sp
ace. In this paper, we introduce it into the computer vision community as a powe
rful metric and an alternative to the widely studied Mahalanobis metric. We show
 that besides its good characteristic in non-Euclidean space, it is a generaliza
tion of Mahalanobis metric in some specific cases. Furthermore, as many Mahalano
bis metric learning, we give two kinds of Cayley-Klein metric learning methods:
MMC Cayley-Klein metric learning and LMNN Cayley-Klein metric learning. Experime
nts have shown the superiority of Cayley-Klein metric over Mahalanobis ones and
the effectiveness of our Cayley-Klein metric learning methods.
****

From Dictionary of Visual Words to Subspaces: Locality-Constrained Affine Subspa
ce Coding
Peihua Li, Xiaoxiao Lu, Qilong Wang; Proceedings of the IEEE Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2015, pp. 2348-2357
The locality-constrained linear coding (LLC) is a very successful feature coding
 method in image classification. It makes known the importance of locality const
raint which brings high efficiency and local smoothness of the codes. However, i
n the LLC method the geometry of feature space is described by an ensemble of re
presentative points (visual words) while discarding the geometric structure imme
diately surrounding them. Such a dictionary only provides a crude, piecewise con
stant approximation of the data manifold. To approach this problem, we propose a
 novel feature coding method called locality-constrained affine subspace coding
(LASC). The data manifold in LASC is characterized by an ensemble of subspaces a

ttached to the representative points (or affine subspaces), which can provide a piecewise linear approximation of the manifold. Given an input descriptor, we find its top-k neighboring subspaces, in which the descriptor is linearly decomposed and weighted to form the first-order LASC vector. Inspired by the success of usage of higher-order information in image classification, we propose the second-order LASC vector based on the Fisher information metric for further performance improvement. We make experiments on challenging benchmarks and experiments have shown the LASC method is very competitive.
****

FPA-CS: Focal Plane Array-Based Compressive Imaging in Short-Wave Infrared
Huaijin Chen, M. Salman Asif, Aswin C. Sankaranarayanan, Ashok Veeraraghavan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2358-2366
Cameras for imaging in short and mid-wave infrared spectra are significantly more expensive than their counterparts in visible imaging. As a result, high-resolution imaging in those spectrum remains beyond the reach of most consumers. Over the last decade, compressive sensing (CS) has emerged as a potential means to realize inexpensive short-wave infrared cameras. One approach for doing this is the single-pixel camera (SPC) where a single detector acquires coded measurements of a high-resolution image. A computational reconstruction algorithm is then used to recover the image from these coded measurements. Unfortunately, the measurement rate of a SPC is insufficient to enable imaging at high spatial and temporal resolutions. We present a focal plane array-based compressive sensing (FPA-CS) architecture that achieves high spatial and temporal resolutions. The idea is to use an array of SPCs that sense in parallel to increase the measurement rate, and consequently, the achievable spatio-temporal resolution of the camera. We develop a proof-of-concept prototype in the short-wave infrared using a sensor with 64 x 64 pixels; the prototype provides a 4096x increase in the measurement rate compared to the SPC and achieves a megapixel resolution at video rate using CS techniques.
****

BOLD - Binary Online Learned Descriptor For Efficient Image Matching
Vassileios Balntas, Lilian Tang, Krystian Mikolajczyk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2367-2375
In this paper we propose a novel approach to generate a binary descriptor optimized for each image patch independently. The approach is inspired by the linear discriminant embedding that simultaneously increases inter and decreases intra class distances. A set of discriminative and uncorrelated binary tests is established from all possible tests in an offline training process. The patch adapted descriptors are then efficiently built online from a subset of tests which lead to lower intra class distances thus a more robust descriptor. A patch descriptor consists of two binary strings where one represents the results of the tests and the other indicates the subset of the patch-related robust tests that are used for calculating a masked Hamming distance. Our experiments on three different benchmarks demonstrate improvements in matching performance, and illustrate that per-patch optimization outperforms global optimization.
****

Defocus Deblurring and Superresolution for Time-of-Flight Depth Cameras
Lei Xiao, Felix Heide, Matthew O'Toole, Andreas Kolb, Matthias B. Hullin, Kyros Kutulakos, Wolfgang Heidrich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2376-2384
Continuous-wave time-of-flight (ToF) cameras show great promise as low-cost depth image sensors in mobile applications. However, they also suffer from several challenges, including limited illumination intensity, which mandates the use of large numerical aperture lenses, and thus results in a shallow depth of field, making it difficult to capture scenes with large variations in depth. Another shortcoming is the limited spatial resolution of currently available ToF sensors. In this paper we analyze the image formation model for blurred ToF images. By directly working with raw sensor measurements but regularizing the recovered depth and amplitude images, we are able to simultaneously deblur and super-resolve the

output of ToF cameras. Our method outperforms existing methods on both synthetic and real datasets. In the future our algorithm should extend easily to cameras that do not follow the cosine model of continuous-wave sensors, as well as to multi-frequency or multi-phase imaging employed in more recent ToF cameras.
****

Burst Deblurring: Removing Camera Shake Through Fourier Burst Accumulation
Mauricio Delbracio, Guillermo Sapiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2385-2393
Numerous recent approaches attempt to remove image blur due to camera shake, either with one or multiple input images, by explicitly solving an inverse and inherently ill-posed deconvolution problem. If the photographer takes a burst of images, a modality available in virtually all modern digital cameras, we show that it is possible to combine them to get a clean sharp version. This is done without explicitly solving any blur estimation and subsequent inverse problem. The proposed algorithm is strikingly simple: it performs a weighted average in the Fourier domain, with weights depending on the Fourier spectrum magnitude. The method's rationale is that camera shake has a random nature and therefore each image in the burst is generally blurred differently. Experiments with real camera data show that the proposed Fourier Burst Accumulation algorithm achieves state-of-the-art results an order of magnitude faster, with simplicity for on-board implementation on camera phones.
****

SOM: Semantic Obviousness Metric for Image Quality Assessment
Peng Zhang, Wengang Zhou, Lei Wu, Houqiang Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2394-2402
Image quality assessment (IQA) tries to estimate human perception based image visual quality in an objective manner. Existing approaches target this problem with or without reference images. For no-reference image quality assessment, there is no given reference image or any knowledge of the distortion type of the image. Previous approaches measure the image quality from signal level rather than semantic analysis. They typically depend on various features to represent local characteristic of an image. In this paper we propose a new no-reference (NR) image quality assessment (IQA) framework based on semantic obviousness. We discover that semantic-level factors affect human perception of image quality. With such observation, we explore semantic obviousness as a metric to perceive objects of an image. We propose to extract two types of features, one to measure the semantic obviousness of the image and the other to discover local characteristic. Then the two kinds of features are combined for image quality estimation. The principles proposed in our approach can also be incorporated with many existing IQA algorithms to boost their performance. We evaluate our approach on the LIVE dataset. Our approach is demonstrated to be superior to the existing NR-IQA algorithms and comparable to the state-of-the-art full-reference IQA (FR-IQA) methods. Cross-dataset experiments show the generalization ability of our approach.
****

DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection
Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2403-2412
In this paper, we propose deformable deep convolutional neural networks for generic object detection. This new deep learning object detection diagram has innovations in multiple aspects. In the proposed new deep architecture, a new deformation constrained pooling (def-pooling) layer models the deformation of object parts with geometric constraint and penalty. A new pre-training strategy is proposed to learn feature representations more suitable for the object detection task and with good generalization capability. By changing the net structures, training strategies, adding and removing some key components in the detection pipeline, a set of models with large diversity are obtained, which significantly improves the effectiveness of model averaging. The proposed approach improves the mean averaged precision obtained by RCNN, which is the state-of-the-art, from 31% to

50.3% on the ILSVRC2014 detection dataset. Detailed component-wise analysis is also provided through extensive experimental evaluation, which provide a global view for people to understand the deep learning object detection pipeline.
****

Efficient Globally Optimal Consensus Maximisation With Tree Search
Tat-Jun Chin, Pulak Purkait, Anders Eriksson, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2413-2421
Maximum consensus is one of the most popular criteria for robust estimation in computer vision. Despite its widespread use, optimising the criterion is still customarily done by randomised sample-and-test techniques, which do not guarantee optimality of the result. Several globally optimal algorithms exist, but they are too slow to challenge the dominance of randomised methods. We aim to change this state of affairs by proposing a very efficient algorithm for global maximisation of consensus. Under the framework of LP-type methods, we show how consensus maximisation for a wide variety of vision tasks can be posed as a tree search problem. This insight leads to a novel algorithm based on A* search. We propose efficient heuristic and support set updating routines that enable A* search to rapidly find globally optimal results. On common estimation problems, our algorithm is several orders of magnitude faster than previous exact methods. Our work identifies a promising solution for globally optimal consensus maximisation.
****

Mind's Eye: A Recurrent Visual Representation for Image Caption Generation
Xinlei Chen, C. Lawrence Zitnick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2422-2431
In this paper we explore the bi-directional mapping between images and their sentence-based descriptions. Critical to our approach is a recurrent neural network that attempts to dynamically build a visual representation of the scene as a caption is being generated or read. The representation automatically learns to remember long-term visual concepts. Our model is capable of both generating novel captions given an image, and reconstructing visual features given an image description. We evaluate our approach on several tasks. These include sentence generation, sentence retrieval and image retrieval. State-of-the-art results are shown for the task of generating novel image descriptions. When compared to human generated captions, our automatically generated captions are equal to or preferred by humans 21.0% of the time. Results are better than or comparable to state-of-the-art results on the image and sentence retrieval tasks for methods using similar visual features.
****

Hierarchical Sparse Coding With Geometric Prior For Visual Geo-Location
Raghuraman Gopalan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2432-2439
We address the problem of estimating location information of an image using principles from automated representation learning. We pursue a hierarchical sparse coding approach that learns features useful in discriminating images across locations, by initializing it with a geometric prior corresponding to transformations between image appearance space and their corresponding location grouping space using the notion of parallel transport on manifolds. We then extend this approach to account for the availability of heterogeneous data modalities such as geo-tags and videos pertaining to different locations, and also study a relatively under-addressed problem of transferring knowledge available from certain locations to infer the grouping of data from novel locations. We evaluate our approach on several standard datasets such as im2gps, San Francisco and MediaEval2010, and obtain state-of-the-art results.
****

P3.5P: Pose Estimation With Unknown Focal Length
Changchang Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2440-2448
It is well known that the problem of camera pose estimation with unknown focal length has 7 degrees of freedom. Since each image point gives 2 constraints, solv

ing this problem requires a minimum of 3.5 image points of 4 known 3D points, where 0.5 means either x or y coordinate of an image point. We refer to this minimal problem as P3.5P. However, the existing methods require 4 full image points to solve the camera pose and focal length. In this paper, we present a general solution to the true minimal P3.5P problem with up to 10 solutions. The remaining image coordinate is then used to filter the candidate solutions, which typically results in a single solution for good data or no solution for outliers. Experiments show the proposed method significantly improves the efficiency over the state of the art methods while maintaining a high accuracy.
****

Joint Vanishing Point Extraction and Tracking
Till Kroeger, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2449-2457
We present a novel vanishing point (VP) detection and tracking algorithm for calibrated monocular image sequences. Previous VP detection and tracking methods usually assume known camera poses for all frames or detect and track separately. We advance the state-of-the-art by combining VP extraction on a Gaussian sphere with recent advances in multi-target tracking on probabilistic occupancy fields. The solution is obtained by solving a Linear Program (LP). This enables the joint detection and tracking of multiple VPs over sequences. Unlike existing works we do not need known camera poses, and at the same time avoid detecting and tracking in separate steps. We also propose an extension to enforce VP orthogonality. We augment an existing video dataset consisting of 48 monocular videos with multiple annotated VPs in 14448 frames for evaluation. Although the method is designed for unknown camera poses, it is also helpful in scenarios with known poses, since a multi-frame approach in VP detection helps to regularize in frames with weak VP line support.
****

Learning a Non-Linear Knowledge Transfer Model for Cross-View Action Recognition
Hossein Rahmani, Ajmal Mian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2458-2466
This paper concerns action recognition from unseen and unknown views. We propose unsupervised learning of a non-linear model that transfers knowledge from multiple views to a canonical view. The proposed Non-linear Knowledge Transfer Model (NKTM) is a deep network, with weight decay and sparsity constraints, which finds a shared high-level virtual path from videos captured from different unknown viewpoints to the same canonical view. The strength of our technique is that we learn a single NKTM for all actions and all camera viewing directions. Thus, NKTM does not require action labels during learning and knowledge of the camera viewpoints during training or testing. NKTM is learned once only from dense trajectories of synthetic points fitted to mocap data and then applied to real video data. Trajectories are coded with a general codebook learned from the same mocap data. NKTM is scalable to new action classes and training data as it does not require re-learning. Experiments on the IXMAS and N-UCLA datasets show that NKTM outperforms existing state-of-the-art methods for cross-view action recognition.
****

Random Tree Walk Toward Instantaneous 3D Human Pose Estimation
Ho Yub Jung, Soochahn Lee, Yong Seok Heo, Il Dong Yun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2467-2474
The availability of accurate depth cameras have made real-time human pose estimation possible; however, there are still demands for faster algorithms on low power processors. This paper introduces 1000 frames per second pose estimation method on a single core CPU. A large computation gain is achieved by random walk sub-sampling. Instead of training trees for pixel-wise classification, a regression tree is trained to estimate the probability distribution to the direction toward the particular joint, relative to the current position. At test time, the direction for the random walk is randomly chosen from a set of representative directions. The new position is found by a constant step toward the direction, and the distribution for next direction is found at the new position. The continual random walk through 3D space will eventually produce an expectation of step positio

ns, which we estimate as the joint position. A regression tree is built separate ly for each joint. The number of random walk steps can be assigned for each join t so that the computation time is consistent regardless of the size of body segm entation. The experiments show that even with large computation gain, the accura cy is higher or comparable to the state-of-the-art pose estimation methods.
****

Deep Hashing for Compact Binary Codes Learning
Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, Jie Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2475-2483
In this paper, we propose a new deep hashing (DH) approach to learn compact bina ry codes for large scale visual search. Unlike most existing binary codes learni ng methods which seek a single linear projection to map each sample into a binar y vector, we develop a deep neural network to seek multiple hierarchical non-lin ear transformations to learn these binary codes, so that the nonlinear relations hip of samples can be well exploited. Our model is learned under three constrain ts at the top layer of the deep network: 1) the loss between the original real-v alued feature descriptor and the learned binary vector is minimized, 2) the bina ry codes distribute evenly on each bit, and 3) different bits are as independent as possible. To further improve the discriminative power of the learned binary codes, we extend DH into supervised DH (SDH) by including one discriminative ter m into the objective function of DH which simultaneously maximizes the inter-cla ss variations and minimizes the intra-class variations of the learned binary cod es. Experimental results show the superiority of the proposed approach over the state-of-the-arts.
****

Completing 3D Object Shape From One Depth Image
Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, Derek Hoie m; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2484-2493
Our goal is to recover a complete 3D model from a depth image of an object. Exis ting approaches rely on user interaction or apply to a limited class of objects, such as chairs. We aim to fully automatically reconstruct a 3D model from any c ategory. We take an exemplar-based approach: retrieve similar objects in a datab ase of 3D models using view-based matching and transfer the symmetries and surfa ces from retrieved models. We investigate completion of 3D models in three cases : novel view (model in database); novel model (models for other objects of the s ame category in database); and novel category (no models from the category in da tabase).
****

Encoding Based Saliency Detection for Videos and Images
Thomas Mauthner, Horst Possegger, Georg Waltner, Horst Bischof; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2494-2502
We present a novel video saliency detection method to support human activity rec ognition and weakly supervised training of activity detection algorithms. Recent research has emphasized the need for analyzing salient information in videos to minimize dataset bias or to supervise weakly labeled training of activity detec tors. In contrast to previous methods we do not rely on training information giv en by either eye-gaze or annotation data, but propose a fully unsupervised algor ithm to find salient regions within videos. In general, we enforce the Gestalt p rinciple of figure-ground segregation for both appearance and motion cues. We in troduce an encoding approach that allows for efficient computation of saliency b y approximating joint feature distributions. We evaluate our approach on several datasets, including challenging scenarios with cluttered background and camera motion, as well as salient object detection in images. Overall, we demonstrate f avorable performance compared to state-of-the-art methods in estimating both gro und-truth eye-gaze and activity annotations.
****

Online Sketching Hashing

Cong Leng, Jiaxiang Wu, Jian Cheng, Xiao Bai, Hanqing Lu; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2503-2 511

Recently, hashing based approximate nearest neighbor (ANN) search has attracted much attention. Extensive new algorithms have been developed and successfully ap plied to different applications. However, two critical problems are rarely menti oned. First, in real-world applications, the data often comes in a streaming fas hion but most of existing hashing methods are batch based models. Second, when t he dataset becomes huge, it is almost impossible to load all the data into memor y to train hashing models. In this paper, we propose a novel approach to handle these two problems simultaneously based on the idea of data sketching. A sketch of one dataset preserves its major characters but with significantly smaller siz e. With a small size sketch, our method can learn hash functions in an online fa shion, while needs rather low computational complexity and storage space. Extens ive experiments on two large scale benchmarks and one synthetic dataset demonstr ate the efficacy of the proposed method.

****

Enriching Object Detection With 2D-3D Registration and Continuous Viewpoint Esti mation

Christopher Bongsoo Choy, Michael Stark, Sam Corbett-Davies, Silvio Savarese; Pr oceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVP R), 2015, pp. 2512-2520

A large body of recent work on object detection has focused on exploiting 3D CAD model databases to improve detection performance. Many of these approaches work by aligning exact 3D models to images using templates generated from renderings of the 3D models at a set of discrete viewpoints. However, the training procedu res for these approaches are computationally expensive and require gigabytes of memory and storage, while the viewpoint discretization hampers pose estimation p erformance. We propose an efficient method for synthesizing templates from 3D m odels that runs on the fly -- that is, it quickly produces detectors for an arbi trary viewpoint of a 3D model without expensive dataset-dependent training or te mplate storage. Given a 3D model and an arbitrary continuous detection viewpoint , our method synthesizes a discriminative template by extracting features from a rendered view of the object and decorrelating spatial dependences among the fea tures. Our decorrelation procedure relies on a gradient-based algorithm that is more numerically stable than standard decomposition-based procedures, and we eff iciently search for candidate detections by computing FFT-based template convolu tions. Due to the speed of our template synthesis procedure, we are able to perf orm joint optimization of scale, translation, continuous rotation, and focal len gth using Metropolis-Hastings algorithm. We provide an efficient GPU implementat ion of our algorithm, and we validate its performance on 3D Object Classes and P ASCAL3D+ datasets.

****

Representing 3D Texture on Mesh Manifolds for Retrieval and Recognition Applicat ions

Naoufel Werghi, Claudio Tortorici, Stefano Berretti, Alberto Del Bimbo; Proceedi ngs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20 15, pp. 2521-2530

In this paper, we present and experiment a novel approach for representing textu re of 3D mesh manifolds using local binary patterns (LBP). Using a recently prop osed framework [37], we compute LBP directly on the mesh surface, either using g eometric or photometric appearance. Compared to its depth-image counterpart, our approach is distinguished by the following features: a) inherits the intrinsic advantages of mesh surface (e.g., preservation of the full geometry); b) does no t require normalization; c) can accommodate partial matching. In addition, it al lows early-level fusion of the geometry and photometric texture modalities. Thro ugh experiments conducted on two application scenarios, namely, 3D texture retri eval and 3D face recognition, we assess the effectiveness of the proposed soluti on with respect to state of the art approaches.

****

Saliency Propagation From Simple to Difficult
Chen Gong, Dacheng Tao, Wei Liu, Stephen J. Maybank, Meng Fang, Keren Fu, Jie Ya
ng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2015, pp. 2531-2539
Saliency propagation has been widely adopted for identifying the most attractive
 object in an image. The propagation sequence generated by existing saliency det
ection methods is governed by the spatial relationships of image regions, i.e.,
the saliency value is transmitted between two adjacent regions. However, for the
 inhomogeneous difficult adjacent regions, such a sequence may incur wrong propa
gations. In this paper, we attempt to manipulate the propagation sequence for op
timizing the propagation quality. Intuitively, we postpone the propagations to d
ifficult regions and meanwhile advance the propagations to less ambiguous simple
 regions. Inspired by the theoretical results in educational psychology, a novel
 propagation algorithm employing the teaching-to-learn and learning-to-teach str
ategies is proposed to explicitly improve the propagation quality. In the teachi
ng-to-learn step, a teacher is designed to arrange the regions from simple to di
fficult and then assign the simplest regions to the learner. In the learning-to-
teach step, the learner delivers its learning confidence to the teacher to assis
t the teacher to choose the subsequent simple regions. Due to the interactions b
etween the teacher and learner, the uncertainty of original difficult regions is
 gradually reduced, yielding manifest salient objects with optimized background
suppression. Extensive experimental results on benchmark saliency datasets demon
strate the superiority of the proposed algorithm over twelve representative sali
ency detectors.
****

Learning an Efficient Model of Hand Shape Variation From Depth Images
Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, Andrew
Fitzgibbon; Proceedings of the IEEE Conference on Computer Vision and Pattern Re
cognition (CVPR), 2015, pp. 2540-2548
We describe how to learn a compact and efficient model of the surface deformatio
n of human hands. The model is built from a set of noisy depth images of a diver
se set of subjects performing different poses with their hands. We represent the
 observed surface using Loop subdivision of a control mesh that is deformed by o
ur learned parametric shape and pose model. The model simultaneously accounts fo
r variation in subject-specific shape and subject-agnostic pose. Specifically, h
and shape is parameterized as a linear combination of a mean mesh in a neutral p
ose with a small number of offset vectors. This mesh is then articulated using s
tandard linear blend skinning (LBS) to generate the control mesh of a subdivisio
n surface. We define an energy that encourages each depth pixel to be explained
by our model, and the use of a smooth subdivision surface allows us to optimize
for all parameters jointly from a rough initialization. The efficacy of our meth
od is demonstrated using both synthetic and real data, where it is shown that ha
nd shape variation can be represented using only a small number of basis directi
ons. We compare with other approaches including PCA and show a substantial impro
vement in the representation power of our model, while maintaining the efficienc
y of a linear shape basis.
****

On the Minimal Problems of Low-Rank Matrix Factorization
Fangyuan Jiang, Magnus Oskarsson, Kalle Astrom; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2549-2557
Low-rank matrix factorization is an essential problem in many areas including co
mputer vision, with applications in e.g. affine structure-from-motion, photomet
ric stereo, and non-rigid structure from motion. However, very little attention
 has been drawn to minimal cases for this problem or to using the minimal config
uration of observations to find the solution. Minimal problems are useful when e
ither outliers are present or the observation matrix is sparse. In this paper,
we first give some theoretical insights on how to generate all the minimal probl
ems of a given size using Laman graph theory. We then propose a new parametrizat
ion and a building-block scheme to solve these minimal problems by extending the
 solution from a small sized minimal problem. We test our solvers on synthetic d

ata as well as real data with outliers or a large portion of missing data and sh ow that our method can handle the cases when other iterative methods, based on c onvex relaxation, fail.
****

Symmetry-Based Text Line Detection in Natural Scenes
Zheng Zhang, Wei Shen, Cong Yao, Xiang Bai; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2558-2567
Recently, a variety of real-world applications have triggered huge demand for te chniques that can extract textual information from natural scenes. Therefore, sc ene text detection and recognition have become active research topics in compute r vision. In this work, we investigate the problem of scene text detection from an alternative perspective and propose a novel algorithm for it. Different from traditional methods, which mainly make use of the properties of single character s or strokes, the proposed algorithm exploits the symmetry property of character groups and allows for direct extraction of text lines from natural images. The experiments on the latest ICDAR benchmarks demonstrate that the proposed algorit hm achieves state-of-the-art performance. Moreover, compared to conventional app roaches, the proposed algorithm shows stronger adaptability to texts in challeng ing scenarios.
****

DevNet: A Deep Event Network for Multimedia Event Detection and Evidence Recount ing
Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, Alex G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2568-2577
In this paper, we focus on complex event detection in internet videos while also providing the key evidences of the detection results. Convolutional Neural Netw orks (CNNs) have achieved promising performance in image classification and acti on recognition tasks. However, it remains an open problem how to use CNNs for vi deo event detection and recounting, mainly due to the complexity and diversity o f video events. In this work, we propose a flexible deep CNN infrastructure, nam ely Deep Event Network (DevNet), that simultaneously detects pre-defined events and provides key spatial temporal evidences. Taking key frames of videos as inpu t, we first detect the event of interest at the video level by aggregating the C NN features of the key frames. The pieces of evidences which recount the detecti on results, are also automatically localized, both temporally and spatially. The challenge is that we only have video level labels, while the key evidences usua lly take place at the frame levels. Based on the intrinsic property of CNNs, we first generate a spatial-temporal saliency map by back passing through DevNet, w hich then can be used to find the key frames which are most indicative to the ev ent, as well as to localize the specific spatial position, usually an object, in the frame of the highly indicative area. Experiments on the large scale TRECVID 2014 MEDTest dataset demonstrate the promising performance of our method, both for event detection and evidence recounting.
****

Learning to Detect Motion Boundaries
Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, Cordelia Schmid; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 , pp. 2578-2586
We propose a learning-based approach for motion boundary detection. Precise loca lization of motion boundaries is essential for the success of optical flow estim ation, as motion boundaries correspond to discontinuities of the optical flow fi eld. The proposed approach allows to predict motion boundaries, using a structur ed random forest trained on the ground-truth of the MPI-Sintel dataset. The rand om forest leverages several cues at the patch level, namely appearance (RGB colo r) and motion cues (optical flow estimated by state-of-the-art algorithms). Expe rimental results show that the proposed approach is both robust and computationa lly efficient. It significantly outperforms state-of-the-art motion-difference a pproaches on the MPI-Sintel and Middlebury datasets. We compare the results obta ined with several state-of-the-art optical flow approaches and study the impact

of the different cues used in the random forest.Furthermore, we introduce a new dataset, the YouTube Motion Boundaries dataset (YMB), that comprises 60 sequences taken from real-world videos with manually annotated motion boundaries. On this dataset, our approach, although trained on MPI-Sintel, also outperforms by a large margin state-of-the-art optical flow algorithms.
****

Improving Object Proposals With Multi-Thresholding Straddling Expansion
Xiaozhi Chen, Huimin Ma, Xiang Wang, Zhichen Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2587-2595
Recent advances in object detection have exploited object proposals to speed up object searching. However, many of existing object proposal generators have strong localization bias or require computationally expensive diversification strategies. In this paper, we present an effective approach to address these issues. We first propose a simple and useful localization bias measure, called superpixel tightness. Based on the characteristics of superpixel tightness distribution, we propose an effective method, namely multi-thresholding straddling expansion (MTSE) to reduce localization bias via fast diversification. Our method is essentially a box refinement process, which is intuitive and beneficial, but seldom exploited before. The greatest benefit of our method is that it can be integrated into any existing model to achieve consistently high recall across various intersection over union thresholds. Experiments on PASCAL VOC dataset demonstrates that our approach improves numerous existing models significantly with little computational overhead.
****

Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel
Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, Greg Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2596-2605
Many visual recognition problems can be approached by counting instances. To determine whether an event is present in a long internet video, one could count how many frames seem to contain the activity. Classifying the activity of a group of people can be done by counting the actions of individual people. Encoding these cardinality relationships can reduce sensitivity to clutter, in the form of irrelevant frames or individuals not involved in a group activity. Learned parameters can encode how many instances tend to occur in a class of interest. To this end, this paper develops a powerful and flexible framework to infer any cardinality relation between latent labels in a multi-instance model. Hard or soft cardinality relations can be encoded to tackle diverse levels of ambiguity. Experiments on tasks such as human activity recognition, video event detection, and video summarization demonstrate the effectiveness of using cardinality relations for improving recognition results.
****

Unconstrained 3D Face Reconstruction
Joseph Roth, Yiying Tong, Xiaoming Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2606-2615
This paper presents an algorithm for unconstrained 3D face reconstruction. The input to our algorithm is an "unconstrained" collection of face images captured under a diverse variation of poses, expressions, and illuminations, without metadata about cameras or timing. The output of our algorithm is a true 3D face surface model represented as a watertight triangulated surface with albedo data or texture information. 3D face reconstruction from a collection of unconstrained 2D images is a long-standing computer vision problem. Motivated by the success of the state-of-the-art method, we developed a novel photometric stereo-based method with two distinct novelties. First, working with a true 3D model allows us to enjoy the benefits of using images from all possible poses, including profiles. Second, by leveraging emerging face alignment techniques and our novel normal field-based Laplace editing, a combination of landmark constraints and photometric stereo-based normals drives our surface reconstruction. Given large photo collections and a ground truth 3D surface, we demonstrate the effectiveness and str

ength of our algorithm both qualitatively and quantitatively.
****
Becoming the Expert - Interactive Multi-Class Machine Teaching
Edward Johns, Oisin Mac Aodha, Gabriel J. Brostow; Proceedings of the IEEE Confe rence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2616-2624
Compared to machines, humans are extremely good at classifying images into categ ories, especially when they possess prior knowledge of the categories at hand. I f this prior information is not available, supervision in the form of teaching i mages is required. To learn categories more quickly, people should see important and representative images first, followed by less important images later - or n ot at all. However, image-importance is individual-specific, i.e. a teaching ima ge is important to a student if it changes their overall ability to discriminate between classes. Further, students keep learning, so while image-importance dep ends on their current knowledge, it also varies with time. In this work we prop ose an Interactive Machine Teaching algorithm that enables a computer to teach c hallenging visual concepts to a human. Our adaptive algorithm chooses, online, w hich labeled images from a teaching set should be shown to the student as they l earn. We show that a teaching strategy that probabilistically models the student 's ability and progress, based on their correct and incorrect answers, produces better `experts'. We present results using real human participants across severa l varied and challenging real-world datasets.
****
Long-Term Recurrent Convolutional Networks for Visual Recognition and Descriptio n
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhas hini Venugopalan, Kate Saenko, Trevor Darrell; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2625-2634
Models comprised of deep convolutional network layers have dominated recent imag e interpretation tasks; we investigate whether models which are also composition al, or "deep", temporally are effective on tasks involving visual sequences or l abel sequences. We develop a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable, and demonstrate the value of these models on benchmark video recognition tasks, image to sentenc e generation problems, and video narration challenges. In contrast to current mo dels which assume a fixed spatio-temporal receptive field or simple temporal ave raging for sequential processing, recurrent convolutional models are "doubly dee p" in that they can be compositional in spatial and temporal "layers". Such mode ls may have advantages when target concepts are complex and/or training data are limited. Learning long-term dependencies is possible when nonlinearities are in corporated into the network state updates. Long-term RNN models are appealing in that they directly can map variable length inputs (i.e. video frames) to variab le length outputs (i.e. natural language text) and can model complex temporal dy namics; yet they can be optimized with backpropagation. Our recurrent long-term models are directly connected to state-of-the-art visual convnet models and can jointly trained, updating temporal dynamics and convolutional perceptual represe ntations simultaneously. Our results show such models have distinct advantages o ver state-of-the-art models for recognition or generation which are separately d efined and/or optimized.
****
Zero-Shot Object Recognition by Semantic Manifold Distance
Zhenyong Fu, Tao Xiang, Elyor Kodirov, Shaogang Gong; Proceedings of the IEEE Co nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2635-2644
Object recognition by zero-shot learning (ZSL) aims to recognise objects without seeing any visual examples by learning knowledge transfer between seen and unse en object classes. This is typically achieved by exploring a semantic embedding space such as attribute space or semantic word vector space. In such a space, bo th seen and unseen class labels, as well as image features can be embedded (proj ected), and the similarity between them can thus be measured directly. Existing works differ in what embedding space is used and how to project the visual data into the semantic embedding space. Yet, they all measure the similarity in the s

pace using a conventional distance metric (e.g. cosine) that does not consider t
he rich intrinsic structure, i.e. semantic manifold, of the semantic categories
in the embedding space. In this paper we propose to model the semantic manifold
in an embedding space using a semantic class label graph. The semantic manifold
structure is used to redefine the distance metric in the semantic embedding spac
e for more effective ZSL. The proposed semantic manifold distance is computed us
ing a novel absorbing Markov chain process (AMP), which has a very efficient clo
sed-form solution. The proposed new model improves upon and seamlessly unifies v
arious existing ZSL algorithms. Extensive experiments on both the large scale Im
ageNet dataset and the widely used Animal with Attribute (AwA) dataset show that
 our model outperforms significantly the state-of-the-arts.
****

Hyper-Class Augmented and Regularized Deep Learning for Fine-Grained Image Class
ification

Saining Xie, Tianbao Yang, Xiaoyu Wang, Yuanqing Lin; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2645-2654

Deep convolutional neural networks (CNN) have seen tremendous success in large-s
cale generic object recognition. In comparison with generic object recognition,
fine-grained image classification (FGIC) is much more challenging because (i) fi
ne-grained labeled data is much more expensive to acquire (usually requiring dom
ain expertise); (ii) there exists large intra-class and small inter-class varian
ce. Most recent work exploiting deep CNN for image recognition with small traini
ng data adopts a simple strategy: pre-train a deep CNN on a large-scale external
 dataset (e.g., ImageNet) and fine-tune on the small-scale target data to fit th
e specific classification task. In this paper, beyond the fine-tuning strategy,
we propose a systematic framework of learning a deep CNN that addresses the chal
lenges from two new perspectives: (i) identifying easily annotated hyper-classes
 inherent in the fine-grained data and acquiring a large number of hyper-class-l
abeled images from readily available external sources (e.g., image search engine
s), and formulating the problem into multi-task learning; (ii) a novel learning
 model by exploiting a regularization between the fine-grained recognition model
 and the hyper-class recognition model. We demonstrate the success of the propos
ed framework on two small-scale fine-grained datasets (Stanford Dogs and Stanfor
d Cars) and on a large-scale car dataset that we collected.
****

Direct Structure Estimation for 3D Reconstruction

Nianjuan Jiang, Daniel Lin, Minh N. Do, Jiangbo Lu; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2655-2663

Most conventional structure-from-motion (SFM) techniques require camera pose est
imation before computing any scene structure. In this work we show that when com
bined with single/multiple homography estimation, the general Euclidean rigidity
 constraint provides a simple formulation for scene structure recovery without e
xplicit camera pose computation. This direct structure estimation (DSE) opens a
new way to design a SFM system that reverses the order of structure and motion e
stimation. We show that this alternative approach works well for recovering scen
e structure and camera poses from sideway motion given planar or general man-mad
e scenes.
****

Global Supervised Descent Method

Xuehan Xiong, Fernando De la Torre; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2015, pp. 2664-2673

Mathematical optimization plays a fundamental role in solving many problems in c
omputer vision (e.g., camera calibration, image alignment, structure from motion
). It is generally accepted that second order descent methods are the most robus
t, fast, and reliable approaches for nonlinear optimization of a general smooth
function. However, in the context of computer vision, second order descent metho
ds have two main drawbacks: 1) the function might not be analytically differenti
able and numerical approximations are impractical, and 2) the Hessian may be lar
ge and not positive definite. Recently, Supervised Descent Method (SDM), a metho
d that learns the "weighted averaged gradients" in a supervised manner has been

proposed to solve these issues. However, SDM is a local algorithm and it is likely to average conflicting gradient directions. This paper proposes Global SDM (GSDM), an extension of SDM that divides the search space into regions of similar gradient directions. GSDM provides a better and more efficient strategy to minimize non-linear least squares functions in computer vision. We illustrate the effectiveness of GSDM in two problems: non-rigid image alignment and extrinsic camera calibration.

****

Robust Camera Location Estimation by Convex Programming

Onur Ozyesil, Amit Singer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2674-2683

3D structure recovery from a collection of 2D images requires the estimation of the camera locations and orientations, i.e. the camera motion. For large, irregular collections of images, existing methods for the location estimation part, which can be formulated as the inverse problem of estimating n locations from noisy measurements of a subset of the pairwise directions, are sensitive to outliers in direction measurements. In this paper, we firstly provide a complete characterization of well-posed instances of the location estimation problem, by presenting its relation to the existing theory of parallel rigidity. For robust estimation of camera locations, we introduce a two-step approach, comprised of a pairwise direction estimation method robust to outliers in point correspondences between image pairs, and a convex program to maintain robustness to outlier directions. In the presence of partially corrupted measurements, we empirically demonstrate that our convex formulation can even recover the locations exactly. Lastly, we demonstrate the utility of our formulations through experiments on Internet photo collections.

****

Practical Robust Two-View Translation Estimation

Johan Fredriksson, Viktor Larsson, Carl Olsson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2684-2690

Outliers pose a problem in all real structure from motion systems. Due to the use of automatic matching methods one has to expect that a (sometimes very large) portion of the detected correspondences can be incorrect. In this paper we propose a method that estimates the relative translation between two cameras and simultaneously maximizes the number of inlier correspondences. Traditionally, outlier removal tasks have been addressed using RANSAC approaches. However, these are random in nature and offer no guarantees of finding a good solution. If the amount of mismatches is large, the approach becomes costly because of the need to evaluate a large number of random samples. In contrast, our approach is based on the branch and bound methodology which guarantees that an optimal solution will be found. While most optimal methods trade speed for optimality, the proposed algorithm has competitive running times on problem sizes well beyond what is common in practice. Experiments on both real and synthetic data show that the method outperforms state-of-the-art alternatives, including RANSAC, in terms of solution quality. In addition, the approach is shown to be faster than RANSAC in settings with a large amount of outliers.

****

Learning From Massive Noisy Labeled Data for Image Classification

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2691-2699

Large-scale supervised datasets are crucial to train convolutional neural networks (CNNs) for various computer vision problems. However, obtaining a massive amount of well-labeled data is usually very expensive and time consuming. In this paper, we introduce a general framework to train CNNs with only a limited number of clean labels and millions of easily obtained noisy labels. We model the relationships between images, class labels and label noises with a probabilistic graphical model and further integrate it into an end-to-end deep learning system. To demonstrate the effectiveness of our approach, we collect a large-scale real-world clothing classification dataset with both noisy and clean labels. Experiment

s on this dataset indicate that our approach can better correct the noisy labels and improves the performance of trained CNNs.
****

KL Divergence Based Agglomerative Clustering for Automated Vitiligo Grading
Mithun Das Gupta, Srinidhi Srinivasa, Madhukara J., Meryl Antony; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2700-2709
In this paper we present a symmetric KL divergence based agglomerative clustering framework to segment multiple levels of depigmentation in Vitiligo images. The proposed framework starts with a simple merge cost based on symmetric KL divergence. We extend the recent body of work related to Bregman divergence based agglomerative clustering and prove that the symmetric KL divergence is an upper-bound for uni-modal Gaussian distributions. This leads to a very simple yet elegant method for bottomup agglomerative clustering. We introduce albedo and reflectance fields as features for the distance computations. We compare against other established methods to bring out possible pros and cons of the proposed method.
****

Robust Saliency Detection via Regularized Random Walks Ranking
Changyang Li, Yuchen Yuan, Weidong Cai, Yong Xia, David Dagan Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2710-2717
In the field of saliency detection, many graph-based algorithms heavily depend on the accuracy of the pre-processed superpixel segmentation, which leads to significant sacrifice of detail information from the input image. In this paper, we propose a novel bottom-up saliency detection approach that takes advantage of both region-based features and image details. To provide more accurate saliency estimations, we first optimize the image boundary selection by the proposed erroneous boundary removal. By taking the image details and region-based estimations into account, we then propose the regularized random walks ranking to formulate pixel-wised saliency maps from the superpixel-based background and foreground saliency estimations. Experiment results on two public datasets indicate the significantly improved accuracy and robustness of the proposed algorithm in comparison with 12 state-of-the-art saliency detection approaches.
****

Weakly Supervised Semantic Segmentation for Social Images
Wei Zhang, Sheng Zeng, Dequan Wang, Xiangyang Xue; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2718-2726
Image semantic segmentation is the task of partitioning image into several regions based on semantic concepts. In this paper, we learn a weakly supervised semantic segmentation model from social images whose labels are not pixel-level but image-level; furthermore, these labels might be noisy. We present a joint conditional random field model leveraging various contexts to address this issue. More specifically, we extract global and local features in multiple scales by convolutional neural network and topic model. Inter-label correlations are captured by visual contextual cues and label co-occurrence statistics. The label consistency between image-level and pixel-level is finally achieved by iterative refinement. Experimental results on two real-world image datasets PASCAL VOC2007 and SIFT-Flow demonstrate that the proposed approach outperforms state-of-the-art weakly supervised methods and even achieves accuracy comparable with fully supervised methods.
****

Image Specificity
Mainak Jas, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2727-2736
For some images, descriptions written by multiple people are consistent with each other. But for other images, descriptions across people vary considerably. In other words, some images are specific - they elicit consistent descriptions from different people - while other images are ambiguous. Applications involving images and text can benefit from an understanding of which images are specific and which ones are ambiguous. For instance, consider text-based image retrieval. If

a query description is moderately similar to the caption (or reference descripti
on) of an ambiguous image, that query may be considered a decent match to the im
age. But if the image is very specific, a moderate similarity between the query
and the reference description may not be sufficient to retrieve the image. In t
his paper, we introduce the notion of image specificity. We present two mechanis
ms to measure specificity given multiple descriptions of an image: an automated
measure and a measure that relies on human judgement. We analyze image specifici
ty with respect to image content and properties to better understand what makes
an image specific. We then train models to automatically predict the specificity
of an image from image features alone without requiring textual descriptions of
the image. Finally, we show that modeling image specificity leads to improvemen
ts in a text-based image retrieval application.
****

A Multi-Plane Block-Coordinate Frank-Wolfe Algorithm for Training Structural SVM
s With a Costly Max-Oracle
Neel Shah, Vladimir Kolmogorov, Christoph H. Lampert; Proceedings of the IEEE Co
nference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2737-2745
Structural support vector machines (SSVMs) are amongst the best performing metho
ds for structured computer vision tasks, such as semantic image segmentation or
human pose estimation. Training SSVMs, however, is computationally costly, becau
se it requires repeated calls to a structured prediction subroutine (called \emp
h{max-oracle}), which has to solve an optimization problem itself, e.g. a graph
cut. In this work, we introduce a new algorithm for SSVM training that is more
efficient than earlier techniques when the max-oracle is computationally expensi
ve, as it is frequently the case in computer vision tasks. The main idea is to (
i) combine the recent stochastic Block-Coordinate Frank-Wolfe algorithm with eff
icient hyperplane caching, and (ii) use an automatic selection rule for deciding
whether to call the exact max-oracle or to rely on an approximate one based on
the cached hyperplanes. We show experimentally that this strategy leads to fast
er convergence towards the optimum with respect to the number of required oracle
calls, and that this also translates into faster convergence with respect to th
e total runtime when the max-oracle is slow compared to the other steps of the a
lgorithm. A C++ implementation is provided at http://www.ist.ac.at/~vnk
****

Web-Scale Training for Face Identification
Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf; Proceedings of the IE
EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2746-
2754
Scaling machine learning methods to very large datasets has attracted considerab
le attention in recent years, thanks to easy access to ubiquitous sensing and da
ta from the web. We study face recognition and show that three distinct properti
es have surprising effects on the transferability of deep convolutional networks
(CNN): (1) The bottleneck of the network serves as an important transfer learni
ng regularizer, and (2) in contrast to the common wisdom, performance saturation
may exist in CNN's (as the number of training samples grows); we propose a solu
tion for alleviating this by replacing the naive random subsampling of the train
ing set with a bootstrapping process. Moreover, (3) we find a link between the r
epresentation norm and the ability to discriminate in a target domain, which she
ds lights on how such networks represent faces. Based on these discoveries, we a
re able to improve face recognition accuracy on the widely used LFW benchmark, b
oth in the verification (1:1) and identification (1:N) protocols, and directly c
ompare, for the first time, with the state of the art Commercially-Off-The-Shelf
system and show a sizable leap in performance.
****

Dynamically Encoded Actions Based on Spacetime Saliency
Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2755-2764
Human actions typically occur over a well localized extent in both space and tim
e. Similarly, as typically captured in video, human actions have small spatiotem
poral support in image space. This paper capitalizes on these observations by we

ighting feature pooling for action recognition over those areas within a video w
here actions are most likely to occur. To enable this operation, we define a nov
el measure of spacetime saliency. The measure relies on two observations regardi
ng foreground motion of human actors: They typically exhibit motion that contras
ts with that of their surrounding region and they are spatially compact. By usin
g the resulting definition of saliency during feature pooling we show that actio
n recognition performance achieves state-of-the-art levels on three widely consi
dered action recognition datasets. Our saliency weighted pooling can be applied
to essentially any locally defined features and encodings thereof. Additionally,
 we demonstrate that inclusion of locally aggregated spatiotemporal energy featu
res, which efficiently result as a by-product of the saliency computation, furth
er boosts performance over reliance on standard action recognition features alon
e.
****

Three Viewpoints Toward Exemplar SVM
Takumi Kobayashi; Proceedings of the IEEE Conference on Computer Vision and Patt
ern Recognition (CVPR), 2015, pp. 2765-2773
In contrast to category-level or cluster-level classifiers, exemplar SVM is succ
essfully applied to classifying (or detecting) a target object as well as transf
erring instance-level annotations. The method, however, is formulated in a highl
y biased classification problem where only one positive sample is contrasted wit
h a substantial number of negative samples, which makes it difficult to properly
 determine the regularization parameters balancing two types of costs derived fr
om positive and negative samples. In this paper, we present two novel viewpoints
 toward exemplar SVM in addition to the original definition. From these proposed
 viewpoints, we can give light on an intrinsic structure of exemplar SVM, reduci
ng two parameters into only one as well as providing clear intuition on the para
meter, in order to free us from exhaustive parameter tuning. We can also clarify
 how the classifier geometrically works so as to produce homogeneous classificat
ion scores of multiple exemplar SVMs which are comparable to each other without
calibration. In addition, we propose a novel feature transformation method based
 on those viewpoints which contributes to general classification tasks. In the e
xperiments on object detection and image classification, the proposed methods re
garding exemplar SVM exhibit favorable performance.
****

Visual Recognition by Learning From Web Data: A Weakly Supervised Domain General
ization Approach
Li Niu, Wen Li, Dong Xu; Proceedings of the IEEE Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2015, pp. 2774-2783
In this work, we formulate a new weakly supervised domain generalization problem
 for the visual recognition task by using loosely labeled web images/videos as t
raining data. Specifically, we aim to address two challenging issues when learni
ng robust classifiers: 1) enhancing the generalization capability of the learnt
classifiers to any unseen target domain; and 2) coping with noise in the labels
of training web images/videos in the source domain. To address the first issue,
we assume the training web images/videos may come from multiple hidden domains w
ith different data distributions. We then extend the multi-class SVM formulation
 to learn one classifier for each class and each latent domain such that multipl
e classifiers from each class can be effectively integrated to achieve better ge
neralization capability. To address the second issue, we partition the training
samples in each class into multiple clusters. By treating each cluster as a "bag
" and the samples in each cluster as "instances", we formulate a new multi-insta
nce learning (MIL) problem for domain generalization by selecting a subset of tr
aining samples from each training bag and simultaneously learning the optimal cl
assifiers based on the selected samples. Moreover, we also extend our newly prop
osed Weakly Supervised Domain Generalization (WSDG) approach by taking advantage
 of the additional textual descriptions that are only available in the training
web images/videos as privileged information. Extensive experiments on four bench
mark datasets demonstrate the effectiveness of our new approaches for visual rec
ognition by learning from web data.

****
Clustering of Static-Adaptive Correspondences for Deformable Object Tracking
Georg Nebehay, Roman Pflugfelder; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2784-2791
We propose a novel method for establishing correspondences on deformable objects for single-target object tracking. The key ingredient is a dissimilarity measure between correspondences that takes into account their geometric compatibility, allowing us to separate inlier correspondences from outliers. We employ both static correspondences from the initial appearance of the object as well as adaptive correspondences from the previous frame to address the stability-plasticity dilemma. The geometric dissimilarity measure enables us to also disambiguate keypoints that are difficult to match. Based on these ideas we build a keypoint-based tracker that outputs rotated bounding boxes. We demonstrate in a rigorous empirical analysis that this tracker outperforms the state of the art on a dataset of 77 sequences.
****
Geo-Semantic Segmentation
Shervin Ardeshir, Kofi Malcolm Collins-Sibley, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2792-2799
The availability of GIS (Geographical Information System) databases for many urban areas, provides a valuable source of information for improving the performance of many computer vision tasks. In this paper, we propose a method which leverages information acquired from GIS databases to perform semantic segmentation of the image alongside with geo-referencing each semantic segment with its address and geo-location. First, the image is segmented into a set of initial super-pixels. Then, by projecting the information from GIS databases, a set of priors are obtained about the approximate location of the semantic entities such as buildings and streets in the image plane. However, there are significant inaccuracies (misalignments) in the projections, mainly due to inaccurate GPS-tags and camera parameters. In order to address this misalignment issue, we perform data fusion such that it improves the segmentation and GIS projections accuracy simultaneously with an iterative approach. At each iteration, the projections are evaluated and weighted in terms of reliability, and then fused with the super-pixel segmentations. First segmentation is performed using random walks, based on the GIS projections. Then the global transformation which best aligns the projections to their corresponding semantic entities is computed and applied to the projections to further align them to the content of the image. The iterative approach continues until the projections and segments are well aligned.
****
Towards Unified Depth and Semantic Prediction From a Single Image
Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2800-2809
Depth estimation and semantic segmentation are two fundamental problems in image understanding. While the two tasks are strongly correlated and mutually beneficial, they are usually solved separately or sequentially. Motivated by the complementary properties of the two tasks, we propose a unified framework for joint depth and semantic prediction. Given an image, we first use a trained Convolutional Neural Network (CNN) to jointly predict a global layout composed of pixel-wise depth values and semantic labels. By allowing for interactions between the depth and semantic information, the joint network provides more accurate depth prediction than a state-of-the-art CNN trained solely for depth prediction [5]. To further obtain fine-level details, the image is decomposed into local segments for region-level depth and semantic prediction under the guidance of global layout. Utilizing the pixel-wise global prediction and region-wise local prediction, we formulate the inference problem in a two-layer Hierarchical Conditional Random Field (HCRF) to produce the final depth and semantic map. As demonstrated in the experiments, our approach effectively leverages the advantages of both tasks and provides the state-of-the-art results.

****
Towards Force Sensing From Vision: Observing Hand-Object Interactions to Infer Manipulation Forces
Tu-Hoa Pham, Abderrahmane Kheddar, Ammar Qammaz, Antonis A. Argyros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2810-2819
We present a novel, non-intrusive approach for estimating contact forces during hand-object interactions relying solely on visual input provided by a single RGB-D camera. We consider a manipulated object with known geometrical and physical properties. First, we rely on model-based visual tracking to estimate the object's pose together with that of the hand manipulating it throughout the motion. Following this, we compute the object's first and second order kinematics using a new class of numerical differentiation operators. The estimated kinematics is then instantly fed into a second-order cone program that returns a minimal force distribution explaining the observed motion. However, humans typically apply more forces than mechanically required when manipulating objects. Thus, we complete our estimation method by learning these excessive forces and their distribution among the fingers in contact. We provide a full validity analysis of the proposed method by evaluating it based on ground truth data from additional sensors such as accelerometers, gyroscopes and pressure sensors. Experimental results show that force sensing from vision (FSV) is indeed feasible.
****
A MRF Shape Prior for Facade Parsing With Occlusions
Mateusz Kozinski, Raghudeep Gadde, Sergey Zagoruyko, Guillaume Obozinski, Renaud Marlet; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2820-2828
We present a new shape prior formalism for segmentation of rectified facade images. It combines the simplicity of split grammars with unprecedented expressive power: the capability of encoding simultaneous alignment in two dimensions, facade occlusions and irregular boundaries between facade elements. Our method simultaneously segments the visible and occluding objects and recovers the structure of the occluded facade. We formulate the task of finding the most likely image segmentation conforming to a prior of the proposed form as a MAP-MRF problem over the standard 4-connected pixel grid with hard constraints on the classes of neighboring pixels, and propose an efficient optimization algorithm for solving it. We demonstrate state of the art results on a number of facade segmentation datasets.
****
Probability Occupancy Maps for Occluded Depth Images
Timur Bagautdinov, Francois Fleuret, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2829-2837
We propose a novel approach to computing the probabilities of presence of multiple and potentially occluding objects in a scene from a single depth map. To this end, we use a generative model that predicts the distribution of depth images that would be produced if the probabilities of presence were known and then to optimize them so that this distribution explains observed evidence as closely as possible. This allows us to exploit very effectively the available evidence and outperform state-of-the-art methods without requiring large amounts of data, or without using the RGB signal that modern RGB-D sensors also provide.
****
Segment Based 3D Object Shape Priors
Rabeeh Karimi Mahabadi, Christian Hane, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2838-2846
Dense 3D reconstruction still remains a hard task for a broad number of object classes which are not sufficiently textured or contain transparent and reflective parts. Shape priors are the tool of choice when the input data itself is not descriptive enough to get a faithful reconstruction. We propose a novel shape prior formulation that splits the object into multiple convex parts. The reconstruct

ion problem is posed as a volumetric multi-label segmentation. Each of the trans itions between labels is penalized with its individual anisotropic smoothness te rm. This powerful formulation allows us to represent a descriptive shape prior. For the object classes used in this paper the individual segments naturally corr espond to different semantic parts of the object. This leads to a semantic segme ntation as a side product of our shape prior formulation. We evaluate our method on several challenging real-world datasets. Our results show that we can resolv e issues such as undesired holes and disconnected parts. Taking into account a segmentation of the free space, we show that we are able to reconstruct concavit ies, such as the interior of a mug.

****

Shape-From-Template in Flatland

Mathias Gallardo, Daniel Pizarro, Adrien Bartoli, Toby Collins; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2847-2854

Shape-from-Template (SfT) is the problem of inferring the shape of a deformable object as observed in an image using a shape template. We call 2DSfT the 'usual' instance of SfT where the shape is a surface embedded in 3D and the image a 2D projection. We introduce 1DSfT, a novel instance of SfT where the shape is a cur ve embedded in 2D and the image a 1D projection. We focus on isometric deformati ons, for which 2DSfT is a well-posed problem, and admits an analytical local sol ution which may be used to initialize nonconvex refinement. Through a complete t heoretical study of 1DSfT with perspective projection, we show that it is relate d to 2DSfT, but may have very different properties: (i) 1DSfT cannot be exactly solved locally and (ii) 1DSfT cannot be solved uniquely, as it has a discrete am ount of at least two solutions. We then propose two convex initialization algori thms, a local analytical one based on infinitesimal planarity and a global one b ased on inextensibility. We show how nonconvex refinement can be implemented whe re, contrarily to current 2DSfT methods, one may enforce isometry exactly using a novel angle-based parameterization. Finally, our method is tested with simulat ed and real data.

****

Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition

Yixin Zhu, Yibiao Zhao, Song Chun Zhu; Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition (CVPR), 2015, pp. 2855-2864

In this paper, we present a new framework - task-oriented modeling, learning and recognition which aims at understanding the underlying functions, physics and c ausality in using objects as "tools". Given a task, such as, cracking a nut or painting a wall, we represent each object, e.g. a hammer or brush, in a generati ve spatio-temporal representation consisting of four components: i) an affordanc e basis to be grasped by hand; ii) a functional basis to act on a target object (the nut), iii) the imagined actions with typical motion trajectories; and iv) t he underlying physical concepts, e.g. force, pressure, etc. In a learning phase, our algorithm observes only one RGB-D video, in which a rational human picks up one object (i.e. tool) among a number of candidates to accomplish the task. Fro m this example, our algorithm learns the essential physical concepts in the task (e.g. forces in cracking nuts). In an inference phase, our algorithm is given a new set of objects (daily objects or stones), and picks the best choice availab le together with the inferred affordance basis, functional basis, imagined human actions (sequence of poses), and the expected physical quantity that it will pr oduce. From this new perspective, any objects can be viewed as a hammer or a sho vel, and object recognition is not merely memorizing typical appearance examples for each category but reasoning the physical mechanisms in various tasks to ach ieve generalization.

****

Deep Roto-Translation Scattering for Object Classification

Edouard Oyallon, Stephane Mallat; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2865-2873

Dictionary learning algorithms or supervised deep convolution networks have cons iderably improved the efficiency of predefined feature representations such as S

IFT. We introduce a deep scattering convolution network, with predefined wavele t filters over spatial and angular variables. This representation brings an imp ortant improvement to results previously obtained with predefined features over object image databases such as Caltech and CIFAR. The resulting accuracy is comp arable to results obtained with unsupervised deep learning and dictionary based representations. This shows that refining image representations by using geomet ric priors is a promising direction to improve image classification and its un derstanding.
****

Non-Rigid Registration of Images With Geometric and Photometric Deformation by U sing Local Affine Fourier-Moment Matching
Hong-Ren Su, Shang-Hong Lai; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2015, pp. 2874-2882
Registration between images taken with different cameras, from different viewpoi nts or under different lighting conditions is a challenging problem. It needs to solve not only the geometric registration problem but also the photometric matc hing problem. In this paper, we propose to estimate the integrated geometric and photometric transformations between two images based on a local affine Fourier-moment matching framework, which is developed to achieve deformable registration . We combine the local Fourier moment constraints with the smoothness constraint s to determine the local affine transforms in a hierarchal block model. Our expe rimental results on registering some real images related by large color and geom etric transformations show the proposed registration algorithm provides superior image registration results compared to the state-of-the-art image registration methods.
****

Detector Discovery in the Wild: Joint Multiple Instance and Representation Learn ing
Judy Hoffman, Deepak Pathak, Trevor Darrell, Kate Saenko; Proceedings of the IEE E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2883-2 891
We develop methods for detector learning which exploit joint training over both weak and strong labels and which transfer learned perceptual representations fro m strongly-labeled auxiliary tasks. Previous methods for weak-label learning oft en learn detector models independently using latent variable optimization, but f ail to share deep representation knowledge across classes and usually require st rong initialization. Other previous methods transfer deep representations from d omains with strong labels to those with only weak labels, but do not optimize ov er individual latent boxes, and thus may miss specific salient structures for a particular category. We propose a model that subsumes these previous approaches , and simultaneously trains a representation and detectors for categories with e ither weak or strong labels present. We provide a novel formulation of a joint m ultiple instance learning method that includes examples from classification-styl e data when available, and also performs domain transfer learning to improve the underlying detector representation. Our model outperforms known methods on Imag eNet-200 detection with weak labels.
****

Deeply Learned Face Representations Are Sparse, Selective, and Robust
Yi Sun, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Comput er Vision and Pattern Recognition (CVPR), 2015, pp. 2892-2900
This paper designs a high-performance deep convolutional network (DeepID2+) for face recognition. It is learned with the identification-verification supervisory signal. By increasing the dimension of hidden representations and adding super vision to early convolutional layers, DeepID2+ achieves new state-of-the-art on LFW and YouTube Faces benchmarks. Through empirical studies, we have discovered three properties of its deep neural activations critical for the high performan ce: sparsity, selectiveness and robustness. (1) It is observed that neural acti vations are moderately sparse. Moderate sparsity maximizes the discriminative po wer of the deep net as well as the distance between images. It is surprising tha t DeepID2+ still can achieve high recognition accuracy even after the neural res

ponses are binarized. (2) Its neurons in higher layers are highly selective to identities and identity-related attributes. We can identify different subsets of neurons which are either constantly excited or inhibited when different identities or attributes are present. Although DeepID2+ is not taught to distinguish attributes during training, it has implicitly learned such high-level concepts. (3) It is much more robust to occlusions, although occlusion patterns are not included in the training set.

****

Unsupervised Visual Alignment With Similarity Graphs

Fatemeh Shokrollahi Yancheshmeh, Ke Chen, Joni-Kristian Kamarainen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2901-2908

Alignment of semantically meaningful visual patterns, such as object classes, is an important pre-processing step for a number of applications such as object detection and image categorization. Considering the expensive manpower spent on the annotation of supervised alignment methods, unsupervised alignment techniques are more favourable especially for large-scale problems. Fine adjustment can be effectively and efficiently achieved with the recent image congealing methods, but they require moderately good initialisation which is largely invalid in practice. It remains as an open problem how to align images of class examples with large view point changes. Feature-based methods can solve the problem to some degree, but require manual selection of a good seed image and omit the fact that two examples of semantical classes can be visually very different (e.g., Harley-Davison and Scooter``motorbikes''). In this work, we adopt the feature basedapproach,but to overcome the aforementioned drawbacks define visual similarity as an assignment problem which is solved by fast approximation and non-linear optimization.From pair-wise image similarities we construct an image graph which is used to step-wise align,``morph'', an image to another by graph traveling. Our method also automatically finds a suitable seed by novel centrality measure which identifies ``similarity hubs'' in the graph. The proposed approach in the unsupervised manner outperforms the state-of-the-art methods with classes from the popular benchmark datasets.

****

Video Anomaly Detection and Localization Using Hierarchical Feature Representation and Gaussian Process Regression

Kai-Wen Cheng, Yie-Tarng Chen, Wen-Hsien Fang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2909-2917

This paper presents a hierarchical framework for detecting local and global anomalies via hierarchical feature representation and Gaussian process regression. While local anomaly is typically detected as a 3D pattern matching problem, we are more interested in global anomaly that involves multiple normal events interacting in an unusual manner such as car accident. To simultaneously detect local and global anomalies, we formulate the extraction of normal interactions from training video as the problem of efficiently finding the frequent geometric relations of the nearby sparse spatio-temporal interest points. A codebook of interaction templates is then constructed and modeled using Gaussian process regression. A novel inference method for computing the likelihood of an observed interaction is also proposed. As such, our model is robust to slight topological deformations and can handle the noise and data unbalance problems in the training data. Simulations show that our system outperforms the main state-of-the-art methods on this topic and achieves at least 80% detection rates based on three challenging datasets.

****

Inferring 3D Layout of Building Facades From a Single Image

Jiyan Pan, Martial Hebert, Takeo Kanade; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2918-2926

In this paper, we propose a novel algorithm that infers the 3D layout of building facades from a single 2D image of an urban scene. Different from existing methods that only yield coarse orientation labels or qualitative block approximations, our algorithm quantitatively reconstructs building facades in 3D space using

a set of planes mutually related by 3D geometric constraints. Each plane is characterized by a continuous orientation vector and a depth distribution. An optimal solution is reached through inter-planar interactions. Due to the quantitative and plane-based nature of our geometric reasoning, our model is more expressive and informative than existing approaches. Experiments show that our method compares competitively with the state of the art on both 2D and 3D measures, while yielding a richer interpretation of the 3D scene behind the image.
****

Evaluation of Output Embeddings for Fine-Grained Image Classification
Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2927-2936
Image classification has advanced significantly in recent years with the availability of large-scale image sets. However, fine-grained classification remains a major challenge due to the annotation cost of large numbers of fine-grained categories. This project shows that compelling classification performance can be achieved on such categories even without labeled training data. Given image and class embeddings, we learn a compatibility function such that matching embeddings are assigned a higher score than mismatching ones; zero-shot classification of an image proceeds by finding the label yielding the highest joint compatibility score. We use state-of-the-art image features and focus on different supervised attributes and unsupervised output embeddings either derived from hierarchies or learned from unlabeled text corpora. We establish a substantially improved state-of-the-art on the Animals with Attributes and Caltech-UCSD Birds datasets. Most encouragingly, we demonstrate that purely unsupervised output embeddings (learned from Wikipedia and improved with fine-grained text) achieve compelling results, even outperforming the previous supervised state-of-the-art. By combining different output embeddings, we further improve results.
****

Virtual View Networks for Object Reconstruction
Joao Carreira, Abhishek Kar, Shubham Tulsiani, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2937-2946
All that structure from motion algorithms "see" are sets of 2D points. We show that these impoverished views of the world can be faked for the purpose of reconstructing objects in challenging settings, such as from a single image, or from a few ones far apart, by recognizing the object and getting help from a collection of images of other objects from the same class. We synthesize virtual views by computing geodesics on novel networks connecting objects with similar viewpoints, and introduce techniques to increase the specificity and robustness of factorization-based object reconstruction in this setting. We report accurate object shape reconstruction from a single image on challenging PASCAL VOC data, which suggests that the current domain of applications of rigid structure-from-motion techniques may be significantly extended.
****

Real-Time Coarse-to-Fine Topologically Preserving Segmentation
Jian Yao, Marko Boben, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2947-2955
In this paper, we tackle the problem of unsupervised segmentation in the form of superpixels. Our main emphasis is on speed and accuracy. We build on [31] to define the problem as a boundary and topology preserving Markov random field. We propose a coarse to fine optimization technique that speeds up inference in terms of the number of updates by an order of magnitude. Our approach is shown to outperform [31] while employing a single iteration. We evaluate and compare our approach to state-of-the-art superpixel algorithms on the BSD and KITTI benchmarks. Our approach significantly outperforms the baselines in the segmentation metrics and achieves the lowest error on the stereo task.
****

Supervised Mid-Level Features for Word Image Representation
Albert Gordo; Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), 2015, pp. 2956-2964
This paper addresses the problem of learning word image representations: given t he cropped image of a word, we are interested in finding a descriptive, robust, and compact fixed-length representation. Machine learning techniques can then be supplied with these representations to produce models useful for word retrieval or recognition tasks. Although many works have focused on the machine learning aspect once a global representation has been produced, little work has been devo ted to the construction of those base image representations: most works use stan dard coding and aggregation techniques directly on top of standard computer visi on features such as SIFT or HOG. We propose to learn local mid-level features s uitable for building word image representations. These features are learnt by le veraging character bounding box annotations on a small set of training images. H owever, contrary to other approaches that use character bounding box information , our approach does not rely on detecting the individual characters explicitly a t testing time. Our local mid-level features can then be aggregated to produce a global word image signature. When pairing these features with the recent word a ttributes framework of Almaz\'an et al., we obtain results comparable with or be tter than the state-of-the-art on matching and recognition tasks using global d escriptors of only 96 dimensions.
****

Learning Lightness From Human Judgement on Relative Reflectance
Takuya Narihira, Michael Maire, Stella X. Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2965-2973
We develop a new approach to inferring lightness, the perceived reflectance of s urfaces, from a single image. Classic methods view this problem from the perspe ctive of intrinsic image decomposition, where an image is separated into reflect ance and shading components. Rather than reason about reflectance and shading t ogether, we learn to directly predict lightness differences between pixels. Lar ge-scale training from human judgement data on relative reflectance, and patch r epresentations built using deep networks, provide the foundation for our model. Benchmarked on the Intrinsic Images in the Wild dataset, our local lightness mo del achieves on-par performance with the state-of-the-art global lightness model , which incorporates multiple shading/reflectance priors and simultaneous reason ing between pairs of pixels in a dense conditional random field formulation.
****

Scene Classification With Semantic Fisher Vectors
Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, Nuno Vasconcelos; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 5, pp. 2974-2983
With the help of a convolutional neural network~(CNN) trained to recognize objec ts, a scene image is represented as a bag of semantics (BoS). This involves clas sifying image patches using the network and considering the class posterior prob ability vectors as locally extracted semantic descriptors. The image BoS is summ arized using a Fisher vector~(FV) embedding that exploits the properties of the space of these descriptors. The resulting representation is referred to as a sem antic Fisher vector. Two implementations of a semantic FV are investigated. Firs t involves modeling the BoS with a Dirichlet Mixture and computing the Fisher gr adients for this model. Due to the difficulty of mixture modeling on a non-Eucli dean probability simplex, this approach is shown to be unsuccessful. A second im plementation is derived using the interpretation of semantic descriptors as para meters of a multinomial distribution. Like the parameters of any exponential fam ily, these can be projected into their natural parameter space. For a CNN, this is shown equivalent to using inputs of its soft-max layer as patch descriptors. A semantic FV is then computed as a Gaussian Mixture FV in the space of these na tural parameters. This representation is shown to outperform other alternatives such as FVs of features from the intermediate CNN layers or a classifier obtaine d by adapting (fine-tuning) the CNN. The proposed FV represents an embedding for object classification probabilities. As an image representation, therefore, it is complementary to the features obtained from a scene classification CNN. A com bination of the two representations is shown to achieve state-of-the-art results

on MIT Indoor scenes and SUN datasets.
****

Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks
Xiao Lin, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2984-2993
Artificial agents today can answer factual questions. But they fall short on questions that require common sense reasoning. Perhaps this is because most existing common sense databases rely on text to learn and represent knowledge. But much of common sense knowledge is unwritten - partly because it tends to not be interesting enough to talk about, and partly because some common sense is unnatural to articulate in text. While unwritten, it is not unseen. In this paper we leverage semantic common sense knowledge learned from images - i.e. visual common sense - in two textual tasks: fill-in-the-blank and visual paraphrasing. We propose to "imagine" the scene behind the text, and leverage visual cues from the "imagined" scenes in addition to textual cues while answering these questions. We imagine the scenes as a visual abstraction. Our approach outperforms a strong text-only baseline on these tasks. Our proposed tasks can serve as benchmarks to quantitatively evaluate progress in solving tasks that go "beyond recognition". Our code and datasets are publicly available.
****

Co-Saliency Detection via Looking Deep and Wide
Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2994-3002
With the goal of effectively identifying common and salient objects in a group of relevant images, co-saliency detection has become essential for many applications such as video foreground extraction, surveillance, image retrieval, and image annotation. In this paper, we propose a unified co-saliency detection framework by introducing two novel insights: 1) looking deep to transfer higher-level representations by using the convolutional neural network with additional adaptive layers could better reflect the properties of the co-salient objects, especially their consistency among the image group; 2) looking wide to take advantage of the visually similar neighbors beyond a certain image group could effectively suppress the influence of the common background regions when formulating the intra-group consistency. In the proposed framework, the wide and deep information are explored for the object proposal windows extracted in each image, and the co-saliency scores are calculated by integrating the intra-image contrast and intra group consistency via a principled Bayesian formulation. Finally the window-level co-saliency scores are converted to the superpixel-level co-saliency maps through a foreground region agreement strategy. Comprehensive experiments on two benchmark datasets have demonstrated the consistent performance gain of the proposed approach.
****

Adopting an Unconstrained Ray Model in Light-Field Cameras for 3D Shape Reconstruction
Filippo Bergamasco, Andrea Albarelli, Luca Cosmo, Andrea Torsello, Emanuele Rodola, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3003-3012
Due to their recent availability as off-the-shelf commercial devices, light-field cameras has gathered increasing attention from both scientific community and industrial operators. However, their composite imaging formation process hinders the ability to exploit the well consolidated stack of calibration methods that are available for traditional cameras. While several efforts have been done to propose practical approaches, most of them still rely on the quasi-pinhole behaviour of the single microlens involved in the capturing process. This results in several drawbacks, ranging from the difficulties in feature detection, due to the reduced size of each microlens, to the need to adopt a model with a relatively small number of parameters. With this paper we propose to embrace a fully non-parametric model for the imaging and we show that it can be properly calibrated with little effort using a dense active target. This process produces a dense set o

f independent rays that cannot be directly used to produce a conventional image. However, they are an ideal tool for 3D reconstruction tasks, since they are highly redundant, very accurate and they cover a wide range of different baselines. The feasibility and convenience of the process and the accuracy of the obtained calibration are comprehensively evaluated through several experiments.
****

Towards 3D Object Detection With Bimodal Deep Boltzmann Machines Over RGBD Imagery
Wei Liu, Rongrong Ji, Shaozi Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3013-3021
Nowadays, detecting objects in 3D scenes like point clouds has become an emerging challenge with various applications. However, it retains as an open problem due to the deficiency of labeling 3D training data. To deploy an accurate detection algorithm typically resorts to investigating both RGB and depth modalities, which have distinct statistics while correlated with each other. Previous research mainly focus on detecting objects using only one modality, which ignores exploiting the cross-modality cues. In this work, we propose a cross-modality deep learning framework based on deep Boltzmann Machines for 3D Scenes object detection. In particular, we demonstrate that by learning cross-modality feature from RGBD data, it is possible to capture their joint information to reinforce detector trainings in individual modalities. In particular, we slide a 3D detection window in the 3D point cloud to match the exemplar shape, which the lack of training data in 3D domain is conquered via (1) We collect 3D CAD models and 2D positive samples from Internet. (2) adopt pretrained R-CNNs [2] to extract raw feature from both RGB and Depth domains. Experiments on RMRC dataset demonstrate that the bimodal based deep feature learning framework helps 3D scene object detection.
****

An Active Search Strategy for Efficient Object Class Detection
Abel Gonzalez-Garcia, Alexander Vezhnevets, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3022-3031
Object class detectors typically apply a window classifier to all the windows in a large set, either in a sliding window manner or using object proposals. In this paper, we develop an active search strategy that sequentially chooses the next window to evaluate based on all the information gathered before. This results in a substantial reduction in the number of classifier evaluations and in a more elegant approach in general. Our search strategy is guided by two forces. First, we exploit context as the statistical relation between the appearance of a window and its location relative to the object, as observed in the training set. This enables to jump across distant regions in the image (e.g. observing a sky region suggests that cars might be far below) and is done efficiently in a Random Forest framework. Second, we exploit the score of the classifier to attract the search to promising areas surrounding a highly scored window, and to keep away from areas near low scored ones. Our search strategy can be applied on top of any classifier as it treats it as a black-box. In experiments with R-CNN on the challenging SUN2012 dataset, our method matches the detection accuracy of evaluating all windows independently, while evaluating 9x fewer windows.
****

Geodesic Exponential Kernels: When Curvature and Linearity Conflict
Aasa Feragen, Francois Lauze, Soren Hauberg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3032-3042
We consider kernel methods on general geodesic metric spaces and provide both negative and positive results. First we show that the common Gaussian kernel can only be generalized to a positive definite kernel on a geodesic metric space if the space is flat. As a result, for data on a Riemannian manifold, the geodesic Gaussian kernel is only positive definite if the Riemannian manifold is Euclidean. This implies that any attempt to design geodesic Gaussian kernels on curved Riemannian manifolds is futile. However, we show that for spaces with conditionally negative definite distances the geodesic Laplacian kernel can be generalized while retaining positive definiteness. This implies that geodesic Laplacian kerne

ls can be generalized to some curved spaces, including spheres and hyperbolic sp
aces. Our theoretical results are verified empirically.
****

Transformation-Invariant Convolutional Jungles
Dmitry Laptev, Joachim M. Buhmann; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2015, pp. 3043-3051
Many Computer Vision problems arise from information processing of data sources
with nuisance variances like scale, orientation, contrast, perspective foreshort
ening or - in medical imaging - staining and local warping. In most cases these
variances can be stated a priori and can be used to improve the generalization o
f recognition algorithms. We propose a novel supervised feature learning approac
h, which efficiently extracts information from these constraints to produce inte
rpretable, transformation-invariant features. The proposed method can incorporat
e a large class of transformations, e.g., shifts, rotations, change of scale, mo
rphological operations, non-linear distortions, photometric transformations, etc
. These features boost the discrimination power of a novel image classification
and segmentation method, which we call Transformation-Invariant Convolutional Ju
ngles (TICJ). We test the algorithm on two benchmarks in face recognition and me
dical imaging, where it achieves state of the art results, while being computati
onally significantly more efficient than Deep Neural Networks.
****

Exemplar SVMs as Visual Feature Encoders
Joaquin Zepeda, Patrick Perez; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2015, pp. 3052-3060
In this work, we investigate the use of exemplar SVMs (linear SVMs trained with
one positive example only and a vast collection of negative examples) as encoder
s that turn generic image features into new, task-tailored features. The propose
d feature encoding leverages the ability of the exemplar-SVM (E-SVM) classifier
to extract, from the original representation of the exemplar image, what is uniq
ue about it. While existing image description pipelines rely on the intuition of
 the designer to encode uniqueness into the feature encoding process, our propos
ed approach does it explicitly relative to a "universe" of features represented
by the generic negatives. We show that such a post-processing enhances the perfo
rmance of state-of-the art image retrieval methods based on aggregated image fea
tures, as well as the performance of nearest class mean and K-nearest neighbor i
mage classification methods. We establish these advantages for several features,
 including "traditional" features as well as features derived from deep convolut
ional neural nets. As an additional contribution, we also propose a recursive ex
tension of this E-SVM encoding scheme (RE-SVM) that provides further performance
 gains.
****

Object Scene Flow for Autonomous Vehicles
Moritz Menze, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2015, pp. 3061-3070
This paper proposes a novel model and dataset for 3D scene flow estimation with
an application to autonomous driving. Taking advantage of the fact that outdoor
scenes often decompose into a small number of independently moving objects, we r
epresent each element in the scene by its rigid motion parameters and each super
pixel by a 3D plane as well as an index to the corresponding object. This minima
l representation increases robustness and leads to a discrete-continuous CRF whe
re the data term decomposes into pairwise potentials between superpixels and obj
ects. Moreover, our model intrinsically segments the scene into its constituting
 dynamic components. We demonstrate the performance of our model on existing ben
chmarks as well as a novel realistic dataset with scene flow ground truth. We ob
tain this dataset by annotating 400 dynamic scenes from the KITTI raw data colle
ction using detailed 3D CAD models for all vehicles in motion. Our experiments a
lso reveal novel challenges which cannot be handled by existing methods.
****

Reflectance Hashing for Material Recognition
Hang Zhang, Kristin Dana, Ko Nishino; Proceedings of the IEEE Conference on Comp

uter Vision and Pattern Recognition (CVPR), 2015, pp. 3071-3080
We introduce a novel method for using reflectance to identify materials. Reflectance offers a unique signature of the material but is challenging to measure and use for recognizing materials due to its high-dimensionality. In this work, one-shot reflectance of a material surface which we refer to as a reflectance disk is capturing using a unique optical camera. The pixel coordinates of these reflectance disks correspond to the surface viewing angles. The reflectance has class-specific stucture and angular gradients computed in this reflectance space reveal the material class. These reflectance disks encode discriminative information for efficient and accurate material recognition. We introduce a framework called reflectance hashing that models the reflectance disks with dictionary learning and binary hashing. We demonstrate the effectiveness of reflectance hashing for material recognition with a number of real-world materials.
****

Joint Photo Stream and Blog Post Summarization and Exploration
Gunhee Kim, Seungwhan Moon, Leonid Sigal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3081-3089
We propose an approach that utilizes large collections of photo streams and blog posts, two of the most prevalent sources of data on the Web, for joint story-based summarization and exploration. Blogs consist of sequences of images and associated text; they portray events and experiences with concise sentences and representative images. We leverage blogs to help achieve story-based semantic summarization of collections of photo streams. In the opposite direction, blog posts can be enhanced with sets of photo streams by showing interpolations between consecutive images in the blogs. We formulate the problem of joint alignment from blogs to photo streams and photo stream summarization in a unified latent ranking SVM framework. We alternate between solving the two coupled latent SVM problems, by first fixing the summarization and solving for the alignment from blog images to photo streams and vice versa. On a newly collected large-scale Disneyland dataset of 10K blogs (120K associated images) and 6K photo streams (540K images), we demonstrate that blog posts and photo streams are mutually beneficial for summarization, exploration, semantic knowledge transfer, and photo interpolation.
****

Video Summarization by Learning Submodular Mixtures of Objectives
Michael Gygli, Helmut Grabner, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3090-3098
We present a novel method for summarizing raw, casually captured videos. The objective is to create a short summary that still conveys the story. It should thus be both, interesting and representative for the input video. Previous methods often used simplified assumptions and only optimized for one of these goals. Alternatively, they used hand-defined objectives that were optimized sequentially by making consecutive hard decisions. This limits their use to a particular setting. Instead, we introduce a new method that (i) uses a supervised approach in order to learn the importance of global characteristics of a summary and (ii) jointly optimizes for multiple objectives and thus creates summaries that posses multiple properties of a good summary. Experiments on two challenging and very diverse datasets demonstrate the effectiveness of our method, where we outperform or match current state-of-the-art.
****

Building Proteins in a Day: Efficient 3D Molecular Reconstruction
Marcus A. Brubaker, Ali Punjani, David J. Fleet; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3099-3108
Discovering the 3D atomic structure of molecules such as proteins and viruses is a fundamental research problem in biology and medicine. Electron Cryomicroscopy (Cryo-EM) is a promising vision-based technique for structure estimation which attempts to reconstruct 3D structures from 2D images. This paper addresses the challenging problem of 3D reconstruction from 2D Cryo-EM images. A new framework for estimation is introduced which relies on modern stochastic optimization techniques to scale to large datasets. We also introduce a novel technique which reduces the cost of evaluating the objective function during optimization by over f

iver orders of magnitude. The net result is an approach capable of estimating 3D molecular structure from large scale datasets in about a day on a single workstation.
****

Learning Descriptors for Object Recognition and 3D Pose Estimation
Paul Wohlhart, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3109-3118
Detecting poorly textured objects and estimating their 3D pose reliably is still a very challenging problem. We introduce a simple but powerful approach to computing descriptors for object views that efficiently capture both the object identity and 3D pose. By contrast with previous manifold-based approaches, we can rely on the Euclidean distance to evaluate the similarity between descriptors, and therefore use scalable Nearest Neighbor search methods to efficiently handle a large number of objects under a large range of poses. To achieve this, we train a Convolutional Neural Network to compute these descriptors by enforcing simple similarity and dissimilarity constraints between the descriptors. We show that our constraints nicely untangle the images from different objects and different views into clusters that are not only well-separated but also structured as the corresponding sets of poses: The Euclidean distance between descriptors is large when the descriptors are from different objects, and directly related to the distance between the poses when the descriptors are from the same object. These important properties allow us to outperform state-of-the-art object views representations on challenging RGB and RGB-D data.
****

Image Partitioning Into Convex Polygons
Liuyun Duan, Florent Lafarge; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3119-3127
The over-segmentation of images into atomic regions has become a standard and powerful tool in Vision. Traditional superpixel methods, that operate at the pixel level, cannot directly capture the geometric information disseminated into the images. We propose an alternative to these methods by operating at the level of geometric shapes. Our algorithm partitions images into convex polygons. It presents several interesting properties in terms of geometric guarantees, region compactness and scalability. The overall strategy consists in building a Voronoi diagram that conforms to preliminarily detected line-segments, before homogenizing the partition by spatial point process distributed over the image gradient. Our method is particularly adapted to images with strong geometric signatures, typically man-made objects and environments. We show the potential of our approach with experiments on large-scale images and comparisons with state-of-the-art superpixel methods.
****

Deep Visual-Semantic Alignments for Generating Image Descriptions
Andrej Karpathy, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3128-3137
We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We demonstrate that our alignment model produces state of the art results in retrieval experiments on Flickr8K, Flickr30K and MSCOCO datasets. We then show that the generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations.
****

Unsupervised Learning of Complex Articulated Kinematic Structures Combining Motion and Skeleton Information

Hyung Jin Chang, Yiannis Demiris; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3138-3146

In this paper we present a novel framework for unsupervised kinematic structure learning of complex articulated objects from a single-view image sequence. In contrast to prior motion information based methods, which estimate relatively simple articulations, our method can generate arbitrarily complex kinematic structures with skeletal topology by a successive iterative merge process. The iterative merge process is guided by a skeleton distance function which is generated from a novel object boundary generation method from sparse points. Our main contributions can be summarised as follows: (i) Unsupervised complex articulated kinematic structure learning by combining motion and skeleton information. (ii) Iterative fine-to-coarse merging strategy for adaptive motion segmentation and structure smoothing. (iii) Skeleton estimation from sparse feature points. (iv) A new highly articulated object dataset containing multi-stage complexity with ground truth. Our experiments show that the proposed method out-performs state-of-the-art methods both quantitatively and qualitatively.

****

Elastic Functional Coding of Human Actions: From Vector-Fields to Latent Variables

Rushil Anirudh, Pavan Turaga, Jingyong Su, Anuj Srivastava; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3147-3155

Human activities observed from visual sensors often give rise to a sequence of smoothly varying features. In many cases, the space of features can be formally defined as a manifold, where the action becomes a trajectory on the manifold. Such trajectories are high dimensional in addition to being non-linear, which can severely limit computations on them. We also argue that by their nature, human actions themselves lie on a much lower dimensional manifold compared to the high dimensional feature space. Learning an accurate low dimensional embedding for actions could have a huge impact in the areas of efficient search and retrieval, visualization, learning, and recognition. Traditional manifold learning addresses this problem for static points in R^n, but its extension to trajectories on Riemannian manifolds is non-trivial and has remained unexplored. The challenge arises due to the inherent non-linearity, and temporal variability that can significantly distort the distance metric between trajectories. To address these issues we use the transport square-root velocity function (TSRVF) space, a recently proposed representation that provides a metric which has favorable theoretical properties such as invariance to group action. We propose to learn the low dimensional embedding with a manifold functional variant of principal component analysis (mfPCA). We show that mfPCA effectively models the manifold trajectories in several applications such as action recognition, clustering and diverse sequence sampling while reducing the dimensionality by a factor of ~250x. The mfPCA features can also be reconstructed back to the original manifold to allow for easy visualization of the latent variable space.

****

Show and Tell: A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU score (the higher the better) on the Pascal dataset is 25, our approach yields 59, to be compar

ed to human performance around 69. We also show BLEU score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, we achieve a BLEU-4 of 27.7, which is the current state-of-the-art.
****

Descriptor Free Visual Indoor Localization With Line Segments
Branislav Micusik, Horst Wildenauer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3165-3173
We present a novel view on the indoor visual localization problem, where we avoid the use of interest points and associated descriptors, which are the basic building blocks of most standard methods. Instead, localization is cast as an alignment problem of the edges of the query image to a 3D model consisting of line segments. The proposed strategy is effective in low-textured indoor environments and in very wide baseline setups as it overcomes the dependency of image descriptors on textures, as well as their limited invariance to view point changes. The basic features of our method, which are prevalent indoors, are line segments. As we will show, they allow for defining an efficient Chamfer distance-based aligning cost, computed through integral contour images, incorporated into a first-best-search strategy. Experiments confirm the efectiveness of the method in terms of both, accuracy and computational complexity.
****

Fixation Bank: Learning to Reweight Fixation Candidates
Jiaping Zhao, Christian Siagian, Laurent Itti; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3174-3182
Predicting where humans will fixate in a scene has many practical applications. Biologically-inspired saliency models decompose visual stimuli into feature maps across multiple scales, and then integrate different feature channels, e.g., in a linear, MAX, or MAP. However, to date there is no universally accepted feature integration mechanism. Here, we propose a new a data-driven solution: We first build a "fixation bank" by mining training samples, which maintains the association between local patterns of activation, in 4 feature channels (color, intensity, orientation, motion) around a given location, and corresponding human fixation density at that location. During testing, we decompose feature maps into blobs, extract local activation patterns around each blob, match those patterns against the fixation bank by group lasso, and determine weights of blobs based on reconstruction errors. Our final saliency map is the weighted sum of all blobs. Our system thus incorporates some amount of spatial and featural context information into the location-dependent weighting mechanism. Tested on two standard data sets (DIEM for training and test, and CRCNS for test only; total of 23,670 training and 15,793 + 4,505 test frames), our model slightly but significantly outperforms 7 state-of-the-art saliency models.
****

Deep Networks for Saliency Detection via Local Estimation and Global Search
Lijun Wang, Huchuan Lu, Xiang Ruan, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3183-3192
This paper presents a saliency detection algorithm by integrating both local estimation and global search. In the local estimation stage, we detect local saliency by using a deep neural network (DNN-L) which learns local patch features to determine the saliency value of each pixel. The estimated local saliency maps are further refined by exploring the high level object concept. In the global search stage, the local saliency map together with global contrast and geometric information are used as global features to describe a set of object candidate regions. Another deep neural network (DNN-G) is trained to predict the saliency score of each object region based on the global features. The final saliency map is generated by a weighted sum of salient object regions. Our method presents two interesting insights. First, local features learned by a supervised scheme can effectively capture local contrast, texture and shape information for saliency detection. Second, complex relationship between different global saliency cues can be captured by deep networks and exploited principally rather than heuristically. Quantitative and qualitative experiments on large benchmark data sets demonstrate that our algorithm performs favorably against the state-of-the-art methods.

****
Reflection Removal Using Ghosting Cues
YiChang Shih, Dilip Krishnan, Fredo Durand, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3193-3201
Photographs taken through glass windows often contain both the desired scene and undesired reflections. Separating the reflection and transmission layers is an important but ill-posed problem that has both aesthetic and practical applications. In this work, we introduce the use of ghosting cues that introduce asymmetry between the layers, thereby helping to significantly reduce the ill-posedness of the problem. Such cues arise from shifted double reflections of the reflected scene off the glass surface. In double-pane windows, each pane reflects shifted and attenuated versions of objects on the same side of the glass as the camera. For single-pane windows, ghosting cues arise from shifted reflections on the two surfaces of the glass pane. Even though the ghosting is sometimes barely perceptible by humans, we can still exploit the cue for layer separation. In this work, we model the ghosted reflection using a double-impulse convolution kernel, and automatically estimate the spatial separation and relative attenuation of the ghosted reflection components. To separate the layers, we propose an algorithm that uses a Gaussian Mixture Model for regularization. Our method is automatic and requires only a single input image. We demonstrate that our approach removes a large fraction of reflections on both synthetic and real-world inputs.
****
A Dataset for Movie Description
Anna Rohrbach, Marcus Rohrbach, Niket Tandon, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3202-3212
Audio Description (AD) provides linguistic descriptions of movies and allows visually impaired people to follow a movie along with their peers. Such descriptions are by design mainly visual and thus naturally form an interesting data source for computer vision and computational linguistics. In this work we propose a novel dataset which contains transcribed ADs, which are temporally aligned to full length HD movies. In addition we also collected the aligned movie scripts which have been used in prior work and compare the two different sources of descriptions. In total the MPII Movie Description dataset (MPII-MD) contains a parallel corpus of over 68K sentences and video snippets from 94 HD movies. We characterize the dataset by benchmarking different approaches for generating video descriptions. Comparing ADs to scripts, we find that ADs are far more visual and describe precisely what is shown rather than what should happen according to the scripts created prior to movie production.
****
Fast and Robust Hand Tracking Using Detection-Guided Optimization
Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3213-3221
Markerless tracking of hands and fingers is a promising enabler for human-computer interaction. However, adoption has been limited because of tracking inaccuracies, incomplete coverage of motions, low framerate, complex camera setups, and high computational requirements. In this paper, we present a fast method for accurately tracking rapid and complex articulations of the hand using a single depth camera. Our algorithm uses a novel detection-guided optimization strategy that increases the robustness and speed of pose estimation. In the detection step, a randomized decision forest classifies pixels into parts of the hand. In the optimization step, a novel objective function combines the detected part labels and a Gaussian mixture representation of the depth to estimate a pose that best fits the depth. Our approach needs comparably less computational resources which makes it extremely fast (50 fps without GPU support). The approach also supports varying static, or moving, camera-to-scene arrangements. We show the benefits of our method by evaluating on public datasets and comparing against previous work.
****

Efficient SDP Inference for Fully-Connected CRFs Based on Low-Rank Decomposition
Peng Wang, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3222-3231
Conditional Random Fields (CRFs) are one of the core technologies in computer vi sion, and have been applied on a wide variety of tasks. Conventional CRFs typica lly define edges between neighboring image pixels, resulting in a sparse graph o ver which inference can be performed efficiently. However, these CRFs fail to mo del more complex priors such as long-range contextual relationships. Fully-conne cted CRFs have thus been proposed. While there are efficient approximate inferen ce methods for such CRFs, usually they are sensitive to initialization and make strong assumptions. In this work, we develop an efficient, yet general SDP algor ithm for inference on fully-connected CRFs. The core of the proposed algorithm i s a tailored quasi-Newton method, which solves a specialized SDP dual problem an d takes advantage of the low-rank matrix approximation for fast computation. Exp eriments demonstrate that our method can be applied to fully-connected CRFs that could not previously be solved, such as those arising in pixel-level image co-s egmentation.
****

Discriminative Learning of Iteration-Wise Priors for Blind Deconvolution
Wangmeng Zuo, Dongwei Ren, Shuhang Gu, Liang Lin, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 323 2-3240
The maximum a posterior (MAP)-based blind deconvolution framework generally invo lves two stages: blur kernel estimation and non-blind restoration. For blur kern el estimation, sharp edge prediction and carefully designed image priors are vit al to the success of MAP. In this paper, we propose a blind deconvolution framew ork together with iteration specific priors for better blur kernel estimation. T he family of hyper-Laplacian $(\Pr (\mathbf{d})\propto {{e}^{-{\left\| \mathbf{ d} \right\|_{p}^{p}}/{\lambda }\;}})$ is adopted for modeling iteration-wise pri ors of image gradients, where each iteration has its own model parameters $\{\la mbda^{(t)}, p^{(t)}\}$. To avoid heavy parameter tuning, all iteration-wise mode l parameters can be learned using our principled discriminative learning model f rom a training set, and can be directly applied to other dataset and real blurry images. Interestingly, with the generalized shrinkage / thresholding operator, negative p value $(p < 0)$ is allowable and we find that it contributes more i n estimating the coarse shape of blur kernel. Experimental results on synthetic and real world images demonstrate that our method achieves better deblurring res ults than the existing gradient prior-based methods. Compared with the state-of- the-art patch prior-based method, our method is competitive in restoration resul ts but is much more efficient.
****

Eye Tracking Assisted Extraction of Attentionally Important Objects From Videos
Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, B.S. Manjunath; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3241-3250
Visual attention is a crucial indicator of the relative importance of objects in visual scenes to human viewers. In this paper, we propose an algorithm to extra ct objects which attract visual attention from videos. As human attention is nat urally biased towards high level semantic objects in visual scenes, this informa tion can be valuable to extract salient objects. The proposed algorithm extracts dominant visual tracks using eye tracking data from multiple subjects on a vide o sequence by a combination of mean-shift clustering and Hungarian algorithm. Th ese visual tracks guide a generic object search algorithm to get candidate objec t locations and extents in every frame. Further, we propose a novel multiple obj ect extraction algorithm by constructing a spatio-temporal mixed graph over obje ct candidates. Bounding box based object extraction inference is performed using binary linear integer programming on a cost function defined over the graph. Fi nally, the object boundaries are refined using grabcut segmentation. The propose d technique outperforms state-of-the-art video segmentation using eye tracking p rior and obtains favorable object extraction over algorithms which do not utiliz

e eye tracking data.
****
Multi-View Feature Engineering and Learning
Jingming Dong, Nikolaos Karianakis, Damek Davis, Joshua Hernandez, Jonathan Balz er, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pa ttern Recognition (CVPR), 2015, pp. 3251-3260
We frame the problem of local representation of imaging data as the computation of minimal sufficient statistics that are invariant to nuisance variability indu ced by viewpoint and illumination. We show that, under very stringent conditions , these are related to "feature descriptors" commonly used in Computer Vision. S uch conditions can be relaxed if multiple views of the same scene are available. We propose a sampling-based and a point-estimate based approximation of such a representation, compared empirically on image-to-(multiple)image matching, for w hich we introduce a multi-view wide-baseline matching benchmark, consisting of a mixture of real and synthetic objects with ground truth camera motion and dense three-dimensional geometry.
****
Self Scaled Regularized Robust Regression
Yin Wang, Caglayan Dicle, Mario Sznaier, Octavia Camps; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3261-326 9
Linear Robust Regression (LRR) seeks to find the parameters of a linear mapping from noisy data corrupted from outliers, such that the number of inliers (i.e. pairs of points where the fitting error of the model is less than a given boun d) is maximized. While this problem is known to be NP hard, several tractable re laxations have been recently proposed along with theoretical conditions guarant eeing exact recovery of the parameters of the model. However, these relaxations may perform poorly in cases where the fitting error for the outliers is large. In addition, these approaches cannot exploit available a-priori information, suc h as co-occurrences. To circumvent these difficulties, in this paper we present an alternative approach to robust regression. Our main result shows that this ap proach is equivalent to a "self-scaled" l-1 regularized robust regression proble m, where the cost function is automatically scaled, with scalings that depend on the a-priori information. Thus, the proposed approach achieves substantially b etter performance than traditional regularized approaches in cases where the ou tliers are far from the linear manifold spanned by the inliers, while at the sam e time exhibits the same theoretical recovery properties. These results are illu strated with several application examples using both synthetic and real data.
****
Simultaneous Feature Learning and Hash Coding With Deep Neural Networks
Hanjiang Lai, Yan Pan, Ye Liu, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3270-3278
Similarity-preserving hashing is a widely-used method for nearest neighbour sear ch in large-scale image retrieval tasks. For most existing hashing methods, an i mage is first encoded as a vector of hand-engineering visual features, followed by another separate projection or quantization step that generates binary codes. However, such visual feature vectors may not be optimally compatible with the c oding process, thus producing sub-optimal hashing codes. In this paper, we propo se a deep architecture for supervised hashing, in which images are mapped into b inary codes via carefully designed deep neural networks. The pipeline of the pro posed deep architecture consists of three building blocks: 1) a sub-network with a stack of convolution layers to produce the effective intermediate image featu res; 2) divide-and-encoding module to divide the intermediate image features int o multiple branches, each encoded into one hash bit; and 3) a triplet ranking lo ss designed to characterize that one image is more similar to the second image t han to the third one. Extensive evaluations on several benchmark image datasets show that the proposed simultaneous feature learning and hash coding pipeline br ings substantial improvements over other state-of-the-art supervised or unsuperv ised hashing methods.
****

MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching
Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, Alexander C. Berg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3279-3286
Motivated by recent successes on learning feature representations and on learning feature comparison functions, we propose a unified approach to combining both for training a patch matching system. Our system, dubbed MatchNet, consists of a deep convolutional network that extracts features from patches and a network of three fully connected layers that computes a similarity between the extracted features. To ensure experimental repeatability, we train MatchNet on standard datasets and employ an input sampler to augment the training set with synthetic exemplar pairs that reduce overfitting. Once trained, we achieve better computational efficiency during matching by disassembling MatchNet and separately applying the feature computation and similarity networks in two sequential stages. We perform a comprehensive set of experiments on standard datasets to carefully study the contributions of each aspect of MatchNet, with direct comparisons to established methods. Our results confirm that our unified approach improves accuracy over previous state-of-the-art results on patch matching datasets, while reducing the storage requirement for descriptors. We make pre-trained MatchNet publicly available.
****
Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset)
Jared Heinly, Johannes L. Schonberger, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3287-3295
We propose a novel, large-scale, structure-from-motion framework that advances the state of the art in data scalability from city-scale modeling (millions of images) to world-scale modeling (several tens of millions of images) using just a single computer. The main enabling technology is the use of a streaming-based framework for connected component discovery. Moreover, our system employs an adaptive, online, iconic image clustering approach based on an augmented bag-of-words representation, in order to balance the goals of registration, comprehensiveness, and data compactness. We demonstrate our proposal by operating on a recent publicly available 100 million image crowd-sourced photo collection containing images geographically distributed throughout the entire world. Results illustrate that our streaming-based approach does not compromise model completeness, but achieves unprecedented levels of efficiency and scalability.
****
Exact Bias Correction and Covariance Estimation for Stereo Vision
Charles Freundlich, Michael Zavlanos, Philippos Mordohai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3296-3304
We present an approach for correcting the bias in 3D reconstruction of points imaged by a calibrated stereo rig. Our analysis is based on the observation that, due to quantization error, a 3D point reconstructed by triangulation essentially represents an entire region in space. The true location of the world point that generated the triangulated point could be anywhere in this region. We argue that the reconstructed point, if it is to represent this region in space without bias, should be located at the centroid of this region, which is not what has been done in the literature. We derive the exact geometry of these regions in space, which we call 3D cells, and we show how they can be viewed as uniform distributions of possible pre-images of the pair of corresponding pixels. By assuming a uniform distribution of points in 3D, as opposed to a uniform distribution of the projections of these 3D points on the images, we arrive at a fast and exact computation of the triangulation bias in each cell. In addition, we derive the exact covariance matrices of the 3D cells. We validate our approach in a variety of simulations ranging from 3D reconstruction to camera localization and relative motion estimation. In all cases, we are able to demonstrate a marked improvement compared to conventional techniques for small disparity values, for which bia

s is significant and the required corrections are large.
****

Computing Similarity Transformations From Only Image Correspondences
Chris Sweeney, Laurent Kneip, Tobias Hollerer, Matthew Turk; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3305-3313

We propose a novel solution for computing the relative pose between two generalized cameras that includes reconciling the internal scale of the generalized cameras. This approach can be used to compute a similarity transformation between two coordinate systems, making it useful for loop closure in visual odometry and registering multiple structure from motion reconstructions together. In contrast to alternative similarity transformation methods, our approach uses 2D-2D image correspondences thus is not subject to the depth uncertainty that often arises with 3D points. We utilize a known vertical direction (which may be easily obtained from IMU data or vertical vanishing point detection) of the generalized cameras to solve the generalized relative pose and scale problem as an efficient Quadratic Eigenvalue Problem. To our knowledge, this is the first method for computing similarity transformations that does not require any 3D information. Our experiments on synthetic and real data demonstrate that this leads to improved performance compared to methods that use 3D-3D or 2D-3D correspondences, especially as the depth of the scene increases.
****

Image Segmentation in Twenty Questions
Christian Rupprecht, Loic Peter, Nassir Navab; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3314-3322

Consider the following scenario between a human user and the computer. Given an image, the user thinks of an object to be segmented within this picture, but is only allowed to provide binary inputs to the computer (yes or no). In these conditions, can the computer guess this hidden segmentation by asking well-chosen questions to the user? We introduce a strategy for the computer to increase the accuracy of its guess in a minimal number of questions. At each turn, the current belief about the answer is encoded in a Bayesian fashion via a probability distribution over the set of all possible segmentations. To efficiently handle this huge space, the distribution is approximated by sampling representative segmentations using an adapted version of the Metropolis-Hastings algorithm, whose proposal moves build on a geodesic distance transform segmentation method. Following a dichotomic search, the question halving the weighted set of samples is finally picked, and the provided answer is used to update the belief for the upcoming rounds. The performance of this strategy is assessed on three publicly available datasets with diverse visual properties. Our approach shows to be a tractable and very adaptive solution to this problem.
****

Interaction Part Mining: A Mid-Level Approach for Fine-Grained Action Recognition
Yang Zhou, Bingbing Ni, Richang Hong, Meng Wang, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3323-3331

Modeling human-object interactions and manipulating motions lies in the heart of fine-grained action recognition. Previous methods heavily rely on explicit detection of the object being interacted, which requires intensive human labour on object annotation. To bypass this constraint and achieve better classification performance, in this work, we propose a novel fine-grained action recognition pipeline by interaction part proposal and discriminative mid-level part mining. Firstly, we generate a large number of candidate object regions using off-the-shelf object proposal tool, e.g., BING. Secondly, these object regions are matched and tracked across frames to form a large spatio-temporal graph based on the appearance matching and the dense motion trajectories through them. We then propose an efficient approximate graph segmentation algorithm to partition and filter the graph into consistent local dense sub-graphs. These sub-graphs, which are spatio-temporal sub-volumes, represent our candidate interaction parts. Finally, we mi

ne discriminative mid-level part detectors from the features computed over the c
andidate interaction parts. Bag-of-detection scores based on a novel Max-N pooli
ng scheme are computed as the action representation for a video sample. We condu
ct extensive experiments on human-object interaction datasets including MPII Coo
king and MSR Daily Activity 3D. The experimental results demonstrate that the pr
oposed framework achieves consistent improvements over the state-of-the-art acti
on recognition accuracies on the benchmarks, without using any object annotation
.
****

Sparse Projections for High-Dimensional Binary Codes
Yan Xia, Kaiming He, Pushmeet Kohli, Jian Sun; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3332-3339
This paper addresses the problem of learning long binary codes from high-dimensi
onal data. We observe that two key challenges arise while learning and using lon
g binary codes: (1) lack of an effective regularizer for the learned high-dimens
ional mapping and (2) high computational cost for computing long codes. In this
paper, we overcome both these problems by introducing a sparsity encouraging reg
ularizer that reduces the effective number of parameters involved in the learned
 projection operator. This regularizer not only reduces overfitting but, due to
the sparse nature of the projection matrix, also leads to a dramatic reduction i
n the computational cost. To evaluate the effectiveness of our method, we analyz
e its performance on the problems of nearest neighbour search, image retrieval a
nd image classification. Experiments on a number of challenging datasets show th
at our method leads to better accuracy than dense projections (ITQ and LSH) with
 the same code lengths, and meanwhile is over an order of magnitude faster. Furt
hermore, our method is also more accurate and faster than other recently propose
d methods for speeding up high-dimensional binary encoding.
****

Hierarchically-Constrained Optical Flow
Ryan Kennedy, Camillo J. Taylor; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2015, pp. 3340-3348
This paper presents a novel approach to solving optical flow problems using a di
screte, tree-structured MRF derived from a hierarchical segmentation of the imag
e. Our method can be used to find globally optimal matching solutions even for p
roblems involving very large motions. Experiments demonstrate that our approach
is competitive on the MPI-Sintel dataset and that it can significantly outperfor
m existing methods on problems involving large motions.
****

The k-Support Norm and Convex Envelopes of Cardinality and Rank
Anders Eriksson, Trung Thanh Pham, Tat-Jun Chin, Ian Reid; Proceedings of the IE
EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3349-
3357
Sparsity, or cardinality, as a tool for feature selection is extremely common in
 a vast number of current computer vision applications. The k-support norm is
a recently proposed norm with the proven property of providing the tightest conv
ex bound on cardinality over the Euclidean norm unit ball. In this paper we pres
ent a re-derivation of this norm, with the hope of shedding further light on thi
s particular surrogate function. In addition, we also present a connection betwe
en the rank operator, the nuclear norm and the k-support norm. Finally, based
on the results established in this re-derivation, we propose a novel algorithm w
ith significantly improved computational efficiency, empirically validated on a
number of different problems, using both synthetic and real world data.
****

Matching Bags of Regions in RGBD images
Hao Jiang; Proceedings of the IEEE Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2015, pp. 3358-3366
We study the new problem of matching regions between a pair of RGBD images given
 a large set of overlapping region proposals. These region proposals do not have
 a tree hierarchy and are treated as bags of regions. Matching RGBD images using
 bags of region candidates with unstructured relations is a challenging combinat

orial problem. We propose a linear formulation, which optimizes the region selec tion and matching simultaneously so that the matched regions have similar color histogram, shape, and small overlaps, the selected regions have a small number a nd overall low concavity, and they tend to cover both of the images. We efficien tly compute the lower bound by solving a sequence of min-cost bipartite matching problems via Lagrangian relaxation and we obtain the global optimum using branc h and bound. Our experiments show that the proposed method is fast, accurate, an d robust against cluttered scenes.
****

Recurrent Convolutional Neural Network for Object Recognition
Ming Liang, Xiaolin Hu; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2015, pp. 3367-3375
In recent years, the convolutional neural network (CNN) has achieved great succe ss in many computer vision tasks. Partially inspired by neuroscience, CNN shares many properties with the visual system of the brain. A prominent difference is that CNN is typically a feed-forward architecture while in the visual system rec urrent connections are abundant. Inspired by this fact, we propose a recurrent C NN (RCNN) for object recognition by incorporating recurrent connections into eac h convolutional layer. Though the input is static, the activities of RCNN units evolve over time so that the activity of each unit is modulated by the activitie s of its neighboring units. This property enhances the ability of the model to i ntegrate the context information, which is important for object recognition. Lik e other recurrent neural networks, unfolding the RCNN through time can result in an arbitrarily deep network with a fixed number of parameters. Furthermore, the unfolded network has multiple paths, which can facilitate the learning process. The model is tested on four benchmark object recognition datasets: CIFAR-10, CI FAR-100, MNIST and SVHN. With fewer trainable parameters, RCNN outperforms the s tate-of-the-art models on all of these datasets. Increasing the number of parame ters leads to even better performance. These results demonstrate the advantage o f the recurrent structure over purely feed-forward structure for object recognit ion.
****

Feedforward Semantic Segmentation With Zoom-Out Features
Mohammadreza Mostajabi, Payman Yadollahpour, Gregory Shakhnarovich; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3376-3385
We introduce a purely feed-forward architecture for semantic segmentation. We ma p small image elements (superpixels) to rich feature representations extracted f rom a sequence of nested regions of increasing extent. These regions are obtaine d by "zooming out" from the superpixel all the way to scene-level resolution. Th is approach exploits statistical structure in the image and in the label space w ithout setting up explicit structured prediction mechanisms, and thus avoids com plex and expensive inference. Instead superpixels are classified by a feedforwar d multilayer network. Our architecture achieves 69.6% average accuracy on the PA SCAL VOC 2012 test set, and 86.1 pixel accuracy on Stanford Background Dataset.
****

The Aperture Problem for Refractive Motion
Tianfan Xue, Hossein Mobahi, Fredo Durand, William T. Freeman; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3 386-3394
When viewed through a small aperture, a moving image provides incomplete informa tion about the local motion. Only the component of motion along the local image gradient is constrained. In an essential part of optical flow algorithms, inform ation must be aggregated from nearby image locations in order to estimate all co mponents of motion. This limitation of local evidence for estimating optical flo w is called ''the aperture problem''. We pose and solve a generalization of the aperture problem for moving refractive elements. We consider a common setup in air flow imaging or telescope observation: a camera is viewing a static backgrou nd, and an unknown refractive elements undergoing unknown motion between them. T hen we are addressing this fundamental question: what does the local image motio

n tell us about the motion of refractive elements? We show that the information gleaned through a local aperture for this case is very different than that for optical flow. In optical flow, the movement of 1D structure already constrains the motion in a certain direction. However, we cannot infer any information about the refractive motion from the movement of 1D structure in the observed sequence, and can only recover one component of the motion from 2D structure. Results on both simulated and real sequences are shown to illustrate our theory.
****

Saliency-Aware Geodesic Video Object Segmentation
Wenguan Wang, Jianbing Shen, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3395-3402
We introduce an unsupervised, geodesic distance based, salient video object segmentation method. Unlike traditional methods, our method incorporates saliency as prior for object via the computation of robust geodesic measurement. We consider two discriminative visual features: spatial edges and temporal motion boundaries as indicators of foreground object locations. We first generate frame-wise spatiotemporal saliency maps using geodesic distance from these indicators. Building on the observation that foreground areas are surrounded by the regions with high spatiotemporal edge values, geodesic distance provides an initial estimation for foreground and background. Then, high-quality saliency results are produced via the geodesic distances to background regions in the subsequent frames. Through the resulting saliency maps, we build global appearance models for foreground and background. By imposing motion continuity, we establish a dynamic location model for each frame. Finally, the spatiotemporal saliency maps, appearance models and dynamic location models are combined into an energy minimization framework to attain both spatially and temporally coherent object segmentation. Extensive quantitative and qualitative experiments on benchmark video dataset demonstrate the superiority of the proposed method over the state-of-the-art algorithms.
****

DEEP-CARVING: Discovering Visual Attributes by Carving Deep Neural Nets
Sukrit Shankar, Vikas K. Garg, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3403-3412
Most of the approaches for discovering visual attributes in images demand significant supervision, which is cumbersome to obtain. In this paper, we aim to discover visual attributes in a weakly supervised setting that is commonly encountered with contemporary image search engines. For instance, given a noun (say forest) and its associated attributes (say dense, sunlit, autumn), search engines can now generate many valid images for any attribute-noun pair (dense forests, autumn forests, etc). However, images for an attribute-noun pair do not contain any information about other attributes (like which forests in the autumn are dense too). Thus, a weakly supervised scenario occurs: each of the M attributes corresponds to a class such that a training image in class m = 1,...,M contains a single label that indicates the presence of the m(th) attribute only. The task is to discover all the attributes present in a test image. Deep Convolutional Neural Networks (CNNs) have enjoyed remarkable success in vision applications recently. However, in a weakly supervised scenario, widely used CNN training procedures do not learn a robust model for predicting multiple attribute labels simultaneously. The primary reason is that the attributes highly co-occur within the training data, and unlike objects, do not generally exist as well-defined spatial boundaries within the image. To ameliorate this limitation, we propose Deep-Carving, a novel training procedure with CNNs, that helps the net efficiently carve itself for the task of multiple attribute prediction. During training, the responses of the feature maps are exploited in an ingenious way to provide the net with multiple pseudo-labels (for training images) for subsequent iterations. The process is repeated periodically after a fixed number of iterations, and enables the net carve itself iteratively for efficiently disentangling features. Additionally, we contribute a noun-adjective pairing inspired Natural Scenes Attributes Dataset to the research community, CAMIT - NSAD, containing a number of co-occurring attributes within a noun category. We describe, in detail, salient aspects of this dataset. Our experiments on CAMIT-NSAD and the SUN Attributes Dataset, w

ith weak supervision, clearly demonstrate that the Deep-Carved CNNs consistently achieve considerable improvement in the precision of attribute prediction over popular baseline methods.
****
Rent3D: Floor-Plan Priors for Monocular Layout Estimation
Chenxi Liu, Alexander G. Schwing, Kaustav Kundu, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3413-3421
The goal of this paper is to enable a 3D "virtual-tour" of an apartment given a small set of monocular images of different rooms, as well as a 2D floor plan. We frame the problem as inference in a Markov Random Field which reasons about the layout of each room and its relative pose (3D rotation and translation) within the full apartment. This gives us accurate camera pose in the apartment for each image. What sets us apart from past work in layout estimation is the use of floor plans as a source of prior knowledge, as well as localization of each image within a bigger space (apartment). In particular, we exploit the floor plan to impose aspect ratio constraints across the layouts of different rooms, as well as to extract semantic information, e.g., the location of windows which are marked in floor plans. We show that this information can significantly help in resolving the challenging room-apartment alignment problem. We also derive an efficient exact inference algorithm which takes only a few ms per apartment. This is due to the fact that we exploit integral geometry as well as our new bounds on the aspect ratio of rooms which allow us to carve the space, significantly reducing the number of physically possible configurations. We demonstrate the effectiveness of our approach on a new dataset which contains over 200 apartments.
****
Learning a Sequential Search for Landmarks
Saurabh Singh, Derek Hoiem, David Forsyth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3422-3430
We propose a general method to find landmarks in images of objects using both appearance and spatial context. This method is applied without changes to two problems: parsing human body layouts, and finding landmarks in images of birds. Our method learns a sequential search for localizing landmarks, iteratively detecting new landmarks given the appearance and contextual information from the already detected ones. The choice of landmark to be added is opportunistic and depends on the image; for example, in one image a head-shoulder group might be expanded to a head-shoulder-hip group but in a different image to a head-shoulder-elbow group. The choice of initial landmark is similarly image dependent. Groups are scored using a learned function, which is used to expand them greedily. Our scoring function is learned from data labelled with landmarks but without any labeling of a detection order. Our method represents a novel spatial model for the kinematics of groups of landmarks, and displays strong performance on two different model problems.
****
Fully Convolutional Networks for Semantic Segmentation
Jonathan Long, Evan Shelhamer, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440
Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation. Our key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. We then define a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Our fully convolutional network achieves state-of-the-art segmentat

ion of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, a
nd SIFT Flow, while inference takes less than one fifth of a second for a typica
l image.
****

Deep Correlation for Matching Images and Text
Fei Yan, Krystian Mikolajczyk; Proceedings of the IEEE Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2015, pp. 3441-3450
This paper addresses the problem of matching images and captions in a joint late
nt space learnt with deep canonical correlation analysis (DCCA). The image and
caption data are represented by the outputs of the vision and text based deep
neural networks. The high dimensionality of the features presents a great challe
nge in terms of memory and speed complexity when used in DCCA framework. We addr
ess these problems by a GPU implementation and propose methods to deal with over
fitting. This makes it possible to evaluate DCCA approach on popular caption-ima
ge matching benchmarks. We compare our approach to other recently proposed techn
iques and present state of the art results on three datasets.
****

Multi-Objective Convolutional Learning for Face Labeling
Sifei Liu, Jimei Yang, Chang Huang, Ming-Hsuan Yang; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3451-3459
This paper formulates face labeling as a conditional random field with unary and
pairwise classifiers. We develop a novel multi-objective learning method that o
ptimizes a single unified deep convolutional network with two distinct non-struc
tured loss functions: one encoding the unary label likelihoods and the other enc
oding the pairwise label dependencies. Moreover, we regularize the network by us
ing a nonparametric prior as new input channels in addition to the RGB image, an
d show that significant performance improvements can be achieved with a much sma
ller network size. Experiments on both the LFW and Helen datasets demonstrate st
ate-of-the-art results of the proposed algorithm, and accurate labeling results
on challenging images can be obtained by the proposed algorithm for real-world a
pplications.
****

Deep Multiple Instance Learning for Image Classification and Auto-Annotation
Jiajun Wu, Yinan Yu, Chang Huang, Kai Yu; Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3460-3469
The recent development in learning deep representations has demonstrated its wid
e applications in traditional vision tasks like classification and detection. Ho
wever, there has been little investigation on how we could build up a deep learn
ing framework in a weakly supervised setting. In this paper, we attempt to model
deep learning in a weakly supervised learning (multiple instance learning) fram
ework. In our setting, each image follows a dual multi-instance assumption, wher
e its object proposals and possible text annotations can be regarded as two inst
ance sets. We thus design effective systems to exploit the MIL property with dee
p learning strategies from the two ends; we also try to jointly learn the relati
onship between object and annotation proposals. We conduct extensive experiments
and prove that our weakly supervised deep learning framework not only achieves
convincing performance in vision tasks including classification and image annota
tion, but also extracts reasonable region-keyword pairs with little supervision,
on both widely used benchmarks like PASCAL VOC and MIT Indoor Scene 67, and als
o a dataset for image- and patch-level annotations.
****

Multi-Instance Object Segmentation With Occlusion Handling
Yi-Ting Chen, Xiaokai Liu, Ming-Hsuan Yang; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3470-3478
We present a multi-instance object segmentation algorithm to tackle occlusions.
As an object is split into two parts by an occluder, it is nearly impossible to
group the two separate regions into an instance by purely bottom-up schemes. To
address this problem, we propose to incorporate top-down category specific reaso
ning and shape prediction through exemplars into an intuitive energy minimizatio
n framework. We perform extensive evaluations of our method on the challenging P

ASCAL VOC 2012 segmentation set. The proposed algorithm achieves favorable resul
ts on the joint detection and segmentation task against the state-of-the-art met
hod both quantitatively and qualitatively.
****

Material Recognition in the Wild With the Materials in Context Database
Sean Bell, Paul Upchurch, Noah Snavely, Kavita Bala; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3479-3487
Recognizing materials in real-world images is a challenging task. Real-world mat
erials have rich surface texture, geometry, lighting conditions, and clutter, wh
ich combine to make the problem particularly difficult. In this paper, we introd
uce a new, large-scale, open dataset of materials in the wild, the Materials in
Context Database (MINC), and combine this dataset with deep learning to achieve
material recognition and segmentation of images in the wild. MINC is an order o
f magnitude larger than previous material databases, while being more diverse an
d well-sampled across its 23 categories. Using MINC, we train convolutional neur
al networks (CNNs) for two tasks: classifying materials from patches, and simult
aneous material recognition and segmentation in full images. For patch-based cla
ssification on MINC we found that the best performing CNN architectures can achi
eve 85.2% mean class accuracy. We convert these trained CNN classifiers into an
efficient fully convolutional framework combined with a fully connected conditio
nal random field (CRF) to predict the material at every pixel in an image, achie
ving 73.1% mean class accuracy. Our experiments demonstrate that having a large,
 well-sampled dataset such as MINC is crucial for real-world material recognitio
n and segmentation.
****

Understanding Pedestrian Behaviors From Stationary Crowd Groups
Shuai Yi, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE Conference on Com
puter Vision and Pattern Recognition (CVPR), 2015, pp. 3488-3496
Pedestrian behavior modeling and analysis is important for crowd scene understan
ding and has various applications in video surveillance. Stationary crowd groups
 are a key factor influencing pedestrian walking patterns but was largely ignore
d in literature. In this paper, a novel model is proposed for pedestrian behavio
r modeling by including stationary crowd groups as a key component. Through infe
rence on the interactions between stationary crowd groups and pedestrians, our m
odel can be used to investigate pedestrian behaviors. The effectiveness of the p
roposed model is demonstrated through multiple applications, including walking p
ath prediction, destination prediction, personality classification, and abnormal
 event detection. To evaluate our model, a large pedestrian walking route datase
t is built. The walking routes of 12, 684 pedestrians from a one-hour crowd surv
eillance video are manually annotated. It will be released to the public and ben
efit future research on pedestrian behavior analysis and crowd scene understandi
ng.
****

Depth From Focus With Your Mobile Phone
Supasorn Suwajanakorn, Carlos Hernandez, Steven M. Seitz; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3497-3
506
While prior depth from focus and defocus techniques operated on laboratory scene
s, we introduce the first depth from focus (DfF) method capable of handling imag
es from mobile phones and other hand-held cameras. Achieving this goal requires
solving a novel uncalibrated DfF problem and aligning the frames to account for
scene parallax. Our approach is demonstrated on a range of challenging cases and
 produces high quality results.
****

Fusion Moves for Correlation Clustering
Thorsten Beier, Fred A. Hamprecht, Jorg H. Kappes; Proceedings of the IEEE Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3507-3516
Correlation clustering, or multicut partitioning, is widely used in image segmen
tation for partitioning an undirected graph or image with positive and negative
edge weights such that the sum of cut edge weights is minimized. Due to its NP-

hardness, exact solvers do not scale and approximative solvers often give unsatisfactory results. We investigate scalable methods for correlation clustering. To this end we define fusion moves for the correlation clustering problem. Our algorithm iteratively fuses the current and a proposed partitioning which monotonously improves the partitioning and maintains a valid partitioning at all times. Furthermore, it scales to larger datasets, gives near optimal solutions, and at the same time shows a good anytime performance.

****

Second-Order Constrained Parametric Proposals and Sequential Search-Based Structured Prediction for Semantic Segmentation in RGB-D Images

Dan Banica, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3517-3526

We focus on the problem of semantic segmentation based on RGB-D data, with emphasis on analyzing cluttered indoor scenes containing many visual categories and instances. Our approach is based on a parametric figure-ground intensity and depth-constrained proposal process that generates spatial layout hypotheses at multiple locations and scales in the image followed by a sequential inference algorithm that produces a complete scene estimate. Our contributions can be summarized as follows: (1) a generalization of parametric max flow figure-ground proposal methodology to take advantage of intensity and depth information, in order to systematically and efficiently generate the breakpoints of an underlying spatial model in polynomial time, (2) new region description methods based on second-order pooling over multiple features constructed using both intensity and depth channels, (3) a principled search-based structured prediction inference and learning process that resolves conflicts in overlapping spatial partitions and selects regions sequentially towards complete scene estimates, and (4) extensive evaluation of the impact of depth, as well as the effectiveness of a large number of descriptors, both pre-designed and automatically obtained using deep learning, in a difficult RGB-D semantic segmentation problem with 92 classes. We report state of the art results in the challenging NYU Depth Dataset V2, extended for the RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, where currently the proposed model ranks first. Moreover, we show that by combining second-order and deep learning features, over 15% relative accuracy improvements can be additionally achieved. In a scene classification benchmark, our methodology further improves the state of the art by 24%.

****

Metric Imitation by Manifold Transfer for Efficient Vision Applications

Dengxin Dai, Till Kroeger, Radu Timofte, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3527-3536

Metric learning has proved very successful. However, human annotations are necessary. In this paper, we propose an unsupervised method, dubbed Metric Imitation (MI), where metrics over one cheap feature (target features, TFs) are learned by imitating the standard metrics over another sophisticated, off-the-shelf feature (source features, SFs) by transferring the view-independent property manifold structures. In particular, MI consists of: 1) quantifying the properties of source metrics as manifold geometry, 2) transferring the manifold from source domain to target domain, and 3) learning a mapping of TFs so that the manifold is approximated as well as possible in the mapped feature domain. MI is useful in at least two scenarios where: 1) TFs are more efficient computationally and in terms of memory than SFs; and 2) SFs contain privileged information, but they are not available during testing. For the former, MI is evaluated on image clustering, category-based image retrieval, and instance-based object retrieval, with three SFs and three TFs. For the latter, MI is tested on the task of example-based image super-resolution, where high-resolution patches are taken as SFs and low-resolution patches as TFs. Experiments show that MI is able to provide good metrics while avoiding expensive data labeling efforts and that it achieves state-of-the-art performance for image super-resolution. In addition, manifold transfer is an interesting direction of transfer learning.

****

The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose

Silvia Zuffi, Michael J. Black; Proceedings of the IEEE Conference on Computer V
ision and Pattern Recognition (CVPR), 2015, pp. 3537-3546
We propose a new 3D model of the human body that is both realistic and part-base
d. The body is represented by a graphical model in which nodes of the graph corr
espond to body parts that can independently translate and rotate in 3D and defor
m to represent different body shapes and to capture pose-dependent shape variati
ons. Pairwise potentials define a "stitching cost" for pulling the limbs apart,
giving rise to the Stitched Puppet (SP) model. Unlike existing realistic 3D bod
y models, the distributed representation facilitates inference by allowing the m
odel to more effectively explore the space of poses, much like existing 2D picto
rial structures models. We infer pose and body shape using a form of particle-ba
sed max-product belief propagation. This gives SP the realism of recent 3D body
models with the computational advantages of part-based models. We apply SP to tw
o challenging problems involving estimating human shape and pose from 3D data. T
he first is the FAUST mesh alignment challenge, where ours is the first method t
o successfully align all 3D meshes with no pose prior. The second involves estim
ating pose and shape from crude visual hull representations of complex body move
ments.
****

Scene Labeling With LSTM Recurrent Neural Networks
Wonmin Byeon, Thomas M. Breuel, Federico Raue, Marcus Liwicki; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3
547-3555
This paper addresses the problem of pixel-level segmentation and classification
of scene images with an entirely learning-based approach using Long Short Term M
emory (LSTM) recurrent neural networks, which are commonly used for sequence cla
ssification. We investigate two-dimensional (2D) LSTM networks for natural scene
 images taking into account the complex spatial dependencies of labels. Prior me
thods generally have required separate classification and image segmentation sta
ges and/or pre- and post-processing. In our approach, classification, segmentati
on, and context integration are all carried out by 2D LSTM networks, allowing te
xture and spatial model parameters to be learned within a single model. The netw
orks efficiently capture local and global contextual information over raw RGB va
lues and adapt well for complex scene images. Our approach, which has a much low
er computational complexity than prior methods, achieved state-of-the-art perfor
mance over the Stanford Background and the SIFT Flow datasets. In fact, if no pr
e- or post-processing is applied, LSTM networks outperform other state-of-the-ar
t approaches. Hence, only with a single-core Central Processing Unit (CPU), the
running time of our approach is equivalent or better than the compared state-of-
the-art approaches which use a Graphics Processing Unit (GPU). Finally, our netw
orks' ability to visualize feature maps from each layer supports the hypothesis
that LSTM networks are overall suited for image processing tasks.
****

FAemb: A Function Approximation-Based Embedding Method for Image Retrieval
Thanh-Toan Do, Quang D. Tran, Ngai-Man Cheung; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3556-3564
The objective of this paper is to design an embedding method mapping local featu
res describing image (e.g. SIFT) to a higher dimensional representation used for
 image retrieval problem. By investigating the relationship between the linear a
pproximation of a nonlinear function in high dimensional space and state-of-the-
art feature representation used in image retrieval, i.e., VLAD, we first introdu
ce a new approach for the approximation. The embedded vectors resulted by the fu
nction approximation process are then aggregated to form a single representation
 used in the image retrieval framework. The evaluation shows that our embedding
method gives a performance boost over the state of the art in image retrieval, a
s demonstrated by our experiments on the standard public image retrieval benchma
rks.
****

Automatically Discovering Local Visual Material Attributes
Gabriel Schwartz, Ko Nishino; Proceedings of the IEEE Conference on Computer Vis

ion and Pattern Recognition (CVPR), 2015, pp. 3565-3573
Shape cues play an important role in computer vision, but shape is not the only information available in images. Materials, such as fabric and plastic, are discernible in images even when shapes, such as those of an object, are not. We argue that it would be ideal to recognize materials without relying on object cues such as shape. This would allow us to use materials as a context for other vision tasks, such as object recognition. Humans are intuitively able to find visual cues that describe materials. While previous frameworks attempt to recognize these cues (as visual material traits), they rely on a fully-supervised set of training image patches. This requirement is not feasible when multiple annotators and large quantities of images are involved. In this paper, we derive a framework that allows us to discover locally-recognizable material attributes from crowdsourced perceptual material distances. We show that the attributes we discover do in fact separate material categories. Our learned attributes exhibit the same desirable properties as material traits, despite the fact that they are discovered using only partial supervision.
****

Depth Image Enhancement Using Local Tangent Plane Approximations
Kiyoshi Matsuo, Yoshimitsu Aoki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3574-3583
This paper describes a depth image enhancement method for consumer RGB-D cameras. Most existing methods use the pixel-coordinates of the aligned color image. Because the image plane generally has no relationship to the measured surfaces, the global coordinate system is not suitable to handle their local geometries. To improve enhancement accuracy, we use local tangent planes as local coordinates for the measured surfaces. Our method is composed of two steps, a calculation of the local tangents and surface reconstruction. To accurately estimate the local tangents, we propose a color heuristic calculation and an orientation correction using their positional relationships. Additionally, we propose a surface reconstruction method by ray-tracing to local tangents. In our method, accurate depth image enhancement is achieved by using the local geometries approximated by the local tangents. We demonstrate the effectiveness of our method using synthetic and real sensor data. Our method has a high completion rate and achieves the lowest errors in noisy cases when compared with existing techniques.
****

Video Co-Summarization: Video Summarization by Visual Co-Occurrence
Wen-Sheng Chu, Yale Song, Alejandro Jaimes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3584-3592
We present video co-summarization, a novel perspective to video summarization that exploits visual co-occurrence across multiple videos. Motivated by the observation that important visual concepts tend to appear repeatedly across videos of the same topic, we propose to summarize a video by finding shots that co-occur most frequently across videos collected using a topic keyword. The main technical challenge is dealing with the sparsity of co-occurring patterns, out of hundreds to possibly thousands of irrelevant shots in videos being considered. To deal with this challenge, we developed a Maximal Biclique Finding (MBF) algorithm that is optimized to find sparsely co-occurring patterns, discarding less co-occurring patterns even if they are dominant in one video. Our algorithm is parallelizable with closed-form updates, thus can easily scale up to handle a large number of videos simultaneously. We demonstrate the effectiveness of our approach on motion capture and self-compiled YouTube datasets. Our results suggest that summaries generated by visual co-occurrence tend to match more closely with human generated summaries, when compared to several popular unsupervised techniques.
****

Watch and Learn: Semi-Supervised Learning for Object Detectors From Video
Ishan Misra, Abhinav Shrivastava, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3593-3602
We present a semi-supervised approach that localizes multiple unknown object instances in long videos. We start with a handful of labeled boxes and iteratively learn and label hundreds of thousands of object instances. We propose criteria f

or reliable object detection and tracking for constraining the semi-supervised l
earning process and minimizing semantic drift. Our approach does not assume exha
ustive labeling of each object instance in any single frame, or any explicit ann
otation of negative data. Working in such a generic setting allow us to tackle m
ultiple object instances in video, many of which are static. In contrast, existi
ng approaches either do not consider multiple object instances per video, or rel
y heavily on the motion of the objects present. The experiments demonstrate the
effectiveness of our approach by evaluating the automatically labeled data on a
variety of metrics like quality, coverage (recall), diversity, and relevance to
training an object detector.
****

Generalized Tensor Total Variation Minimization for Visual Data Recovery
Xiaojie Guo, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2015, pp. 3603-3611
In this paper, we propose a definition of Generalized Tensor Total Variation nor
m (GTV) that considers both the inhomogeneity and the multi-directionality of re
sponses to derivative-like filters. More specifically, the inhomogeneity simulta
neously preserves high-frequency signals and suppresses noises, while the multi-
directionality ensures that, for an entry in a tensor, more information from its
 neighbors is taken into account. To effectively and efficiently seek the soluti
on of the GTV minimization problem, we design a novel Augmented Lagrange Multipl
ier based algorithm, the convergence of which is theoretically guaranteed. Exper
iments are conducted to demonstrate the superior performance of our method over
the state of the art alternatives on classic visual data recovery applications i
ncluding completion and denoising.
****

Active Learning for Structured Probabilistic Models With Histogram Approximation
Qing Sun, Ankit Laddha, Dhruv Batra; Proceedings of the IEEE Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2015, pp. 3612-3621
Abstract. This paper studies active learning in structured probabilistic models
 such as Conditional Random Fields (CRFs). This is a challenging problem because
 unlike unstructured prediction problems such as binary or multi-class classific
ation, structured prediction problems involve a distribution with an exponential
ly-large support, for instance, over the space of all possible segmentations of
an image. Thus, the entropy of such models is typically intractable to compute.
We propose a crude yet surprisingly effective histogram approximation to the Gib
bs distribution, which replaces the exponentially-large support with a coarsened
 distribution that may be viewed as a histogram over M bins. We show that our ap
proach outperforms a number of baselines and results in a 90%-reduction in the n
umber of annotations needed to achieve nearly the same accuracy as learning from
 the entire dataset.
****

Image Parsing With a Wide Range of Classes and Scene-Level Context
Marian George; Proceedings of the IEEE Conference on Computer Vision and Pattern
 Recognition (CVPR), 2015, pp. 3622-3630
This paper presents a nonparametric scene parsing approach that improves the ove
rall accuracy, as well as the coverage of foreground classes in scene images. We
 first improve the label likelihood estimates at superpixels by merging likeliho
od scores from different probabilistic classifiers. This boosts the classificati
on performance and enriches the representation of less-represented classes. Our
second contribution consists of incorporating semantic context in the parsing pr
ocess through global label costs. Our method does not rely on image retrieval se
ts but rather assigns a global likelihood estimate to each label, which is plugg
ed into the overall energy function. We evaluate our system on two large-scale d
atasets, SIFTflow and LMSun. We achieve state-of-the-art performance on the SIFT
flow dataset and near-record results on LMSun.
****

Bayesian Sparse Representation for Hyperspectral Image Super Resolution
Naveed Akhtar, Faisal Shafait, Ajmal Mian; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3631-3640

Despite the proven efficacy of hyperspectral imaging in many computer vision tas
ks, its widespread use is hindered by its low spatial resolution, resulting from
 hardware limitations. We propose a hyperspectral image super resolution approac
h that fuses a high resolution image with the low resolution hyperspectral image
 using non-parametric Bayesian sparse representation. The proposed approach firs
t infers probability distributions for the material spectra in the scene and the
ir proportions. The distributions are then used to compute sparse codes of the h
igh resolution image. To that end, we propose a generic Bayesian sparse coding s
trategy to be used with Bayesian dictionaries learned with the Beta process. We
theoretically analyze the proposed strategy for its accurate performance. The co
mputed codes are used with the estimated scene spectra to construct the super re
solution hyperspectral image. Exhaustive experiments on two public databases of
ground based hyperspectral images and a remotely sensed image show that the prop
osed approach outperforms the existing state of the art.
****

Semantic Object Segmentation via Detection in Weakly Labeled Video
Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, Changqun Xia; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3641-364
9
Semantic object segmentation in video is an important step for large-scale multi
media analysis. In many cases, however, semantic objects are only tagged at vide
o-level, making them difficult to be located and segmented. To address this prob
lem, this paper proposes an approach to segment semantic objects in weakly label
ed video via object detection. In our approach, a novel video segmentationby-det
ection framework is proposed, which first incorporates object and region detecto
rs pre-trained on still images to generate a set of detection and segmentation p
roposals. Based on the noisy proposals, several object tracks are then initializ
ed by solving a joint binary optimization problem with min-cost flow. As such tr
acks actually provide rough configurations of semantic objects, we thus refine t
he object segmentation while preserving the spatiotemporal consistency by inferr
ing the shape likelihoods of pixels from the statistical information of tracks.
Experimental results on Youtube-Objects dataset and SegTrack v2 dataset demonstr
ate that our method outperforms state-of-the-arts and shows impressive results.
****

Learning With Dataset Bias in Latent Subcategory Models
Dimitris Stamos, Samuele Martelli, Moin Nabi, Andrew McDonald, Vittorio Murino,
Massimiliano Pontil; Proceedings of the IEEE Conference on Computer Vision and P
attern Recognition (CVPR), 2015, pp. 3650-3658
Latent subcategory models (LSMs) offer significant improvements over training fl
at classifiers such as linear SVMs. Training LSMs is a challenging task due to t
he potentially large number of local optima in the objective function and the in
creased model complexity which requires large training set sizes. Often larger d
atasets are available as a collection of heterogeneous datasets. However, previo
us work has highlighted the possible danger of simply training a model from the
combined datasets, due to the presence of bias. In this paper, we present a mode
l which jointly learns an LSM for each dataset as well as a compound LSM. The me
thod provides a means to borrow statistical strength from the datasets while red
ucing their inherent bias. In experiments we demonstrate that the compound LSM,
when tested on Pascal, LabelMe, Caltech101 and SUN in a leave-one-dataset-out fa
shion, achieves an average improvement of over 6.5% over a previous SVM-based un
doing bias approach and an average improvement of over 8.6% over a standard LSM
trained on the concatenation of the datasets. Hence our method provides the best
 of both worlds.
****

Project-Out Cascaded Regression With an Application to Face Alignment
Georgios Tzimiropoulos; Proceedings of the IEEE Conference on Computer Vision an
d Pattern Recognition (CVPR), 2015, pp. 3659-3667
Cascaded regression approaches have been recently shown to achieve state-of-the-
art performance for many computer vision tasks. Beyond its connection to boostin
g, cascaded regression has been interpreted as a learning-based approach to iter

ative optimization methods like the Newton's method. However, in prior work, the connection to optimization theory is limited only in learning a mapping from image features to problem parameters. In this paper, we consider the problem of facial deformable model fitting using cascaded regression and make the following contributions: (a) We propose regression to learn a sequence of averaged Jacobian and Hessian matrices from data, and from them descent directions in a fashion inspired by Gauss-Newton optimization. (b) We show that the optimization problem in hand has structure and devise a learning strategy for a cascaded regression approach that takes the problem structure into account. By doing so, the proposed method learns and employs a sequence of averaged Jacobians and descent directions in a subspace orthogonal to the facial appearance variation; hence, we call it Project-Out Cascaded Regression (PO-CR). (c) Based on the principles of PO-CR, we built a face alignment system that produces remarkably accurate results on the challenging iBUG data set outperforming previously proposed systems by a large margin. Code for our system is available from http://www.cs.nott.ac.uk/~yzt/.
****

Image Retrieval Using Scene Graphs
Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3668-3678
This paper develops a novel framework for semantic image retrieval based on the notion of a scene graph. Our scene graphs represent objects ("man", "boat"), attributes of objects ("boat is white") and relationships between objects ("man standing on boat"). We use these scene graphs as queries to retrieve semantically related images. To this end, we design a conditional random field model that reasons about possible groundings of scene graphs to test images. The likelihoods of these groundings are used as ranking scores for retrieval. We introduce a novel dataset of 5,000 human-generated scene graphs grounded to images and use this dataset to evaluate our method for image retrieval. In particular, we evaluate retrieval using full scene graphs and small scene subgraphs, and show that our method outperforms retrieval methods that use only objects or low-level image features. In addition, we show that our full model can be used to improve object localization compared to baseline methods.
****

Unifying Holistic and Parts-Based Deformable Model Fitting
Joan Alabort-i-Medina, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3679-3688
The construction and fitting of deformable models that capture the degrees of freedom of articulated objects is one of the most popular areas of research in computer vision. The two main approaches are: Holistic Deformable Models (HDMs), which try to represent the object as a whole, and Parts-Based Deformable Models (PBDMs), which model object parts independently. Both models have their own advantages. In this paper we try to marry the previous two frameworks into a unified one that potentially combines the advantages of both. We do so by merging the popular Active Appearance Models (holistic) and Constrained Local Models (part-based) using a novel probabilistic formulation of the fitting problem. To the best of our knowledge, this is the first time that such an idea has been proposed. We show that our unified holistic and part-based formulation achieves state-of-the-art results in the problem of face alignment in-the-wild. Finally, in order to encourage open research and facilitate future comparisons with the proposed method, our code will be made publicly available to the research community.
****

Small Instance Detection by Integer Programming on Object Density Maps
Zheng Ma, Lei Yu, Antoni B. Chan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3689-3697
We propose a novel object detection framework for partially-occluded small instances, such as pedestrians in low resolution surveillance video, cells under a microscope, flocks of small animals (e.g. birds, fishes), or even tiny insects like honeybees and flies. These scenarios are very challenging for traditional detectors, which are typically trained on individual instances. In our approach, we

first estimate the object density map of the input image, and then divide it int
o local regions. For each region, a sliding window (ROI) is passed over the den
sity map to calculate the instance count within each ROI. 2D integer programming
 is used to recover the locations of object instances from the set of ROI counts
, and the global count estimate of the density map is used as a constraint to re
gularize the detection performance. Finally, the bounding box for each instance
is estimated using the local density map. Compared with current small-instance d
etection methods, our proposed approach achieves state-of-the-art performance on
 several challenging datasets including fluorescence microscopy cell images, UCS
D pedestrians, small animals and insects.
****

Motion Part Regularization: Improving Action Recognition via Trajectory Selectio
n

Bingbing Ni, Pierre Moulin, Xiaokang Yang, Shuicheng Yan; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3698-3
706

Dense local motion features such as dense trajectories have been widely used in
action recognition. For most actions, only a few local features (e.g., critical
movements of the hand, arm, leg etc.) are responsible to the action label. There
fore, discovering important motion part will lead to a more discriminative and c
ompact action representation. Inspired by the recent advance in sentence regular
ization for text classification, we introduce a Motion Part Regularization frame
work to mining discriminative semi-local groups of dense trajectories. First, mo
tion part candidates are generated by spatio-temporal grouping of densely sample
d trajectories. Then, we develop a learning objective function which encourages
sparse selection for these trajectory groups in conjunction with a discriminativ
e term. We propose an alternative optimization algorithm to efficiently solve th
is objective function by introducing a set of auxiliary variables. The learned t
rajectory group weights are further utilized for weighted bag-of-feature represe
ntation for unknown action samples. The proposed motion part regularization fram
ework achieves the state-of-the-art performances on several action recognition b
enchmarks.
****

Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection

Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, Jiebo Luo; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3707-371
5

Given the tremendous growth of online videos, video thumbnail, as the common vis
ualization form of video content, is becoming increasingly important to influenc
e user's browsing and searching experience. However, conventional methods for vi
deo thumbnail selection often fail to produce satisfying results as they ignore
the side semantic information (e.g., title, description, and query) associated w
ith the video. As a result, the selected thumbnail cannot always represent video
 semantics and the click-through rate is adversely affected even when the retrie
ved videos are relevant. In this paper, we have developed a multi-task deep visu
al-semantic embedding model, which can automatically select query-dependent vide
o thumbnails according to both visual and side information. Different from most
existing methods, the proposed approach employs the deep visual-semantic embeddi
ng model to directly compute the similarity between the query and video thumbnai
ls by mapping them into a common latent semantic space, where even unseen query-
thumbnail pairs can be correctly matched. In particular, we train the embedding
model by exploring the large-scale and freely accessible click-through video and
 image data, as well as employing a multi-task learning strategy to holistically
 exploit the query-thumbnail relevance from these two highly related datasets. F
inally, a thumbnail is selected by fusing both the representative and query rele
vance scores. The evaluations on 1,000 query-thumbnail dataset labeled by 191 wo
rkers in Amazon Mechanical Turk have demonstrated the effectiveness of our propo
sed method.
****

Fine-Grained Visual Categorization via Multi-Stage Metric Learning

Qi Qian, Rong Jin, Shenghuo Zhu, Yuanqing Lin; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3716-3724
Fine-grained visual categorization (FGVC) is to categorize objects into subordin ate classes instead of basic classes. One major challenge in FGVC is the co-occu rrence of two issues: 1) many subordinate classes are highly correlated and are difficult to distinguish, and 2) there exists the large intra-class variation (e .g., due to object pose). This paper proposes to explicitly address the above tw o issues via distance metric learning (DML). DML addresses the first issue by le arning an embedding so that data points from the same class will be pulled toget her while those from different classes should be pushed apart from each other; a nd it addresses the second issue by allowing the flexibility that only a portion of the neighbors (not all data points) from the same class need to be pulled to gether. However, feature representation of an image is often high dimensional, a nd DML is known to have difficulty in dealing with high dimensional feature vect ors since it would require $O(d^2)$ for storage and $O(d^3)$ for optimization. T o this end, we proposed a multi-stage metric learning framework that divides the large-scale high dimensional learning problem to a series of simple subproblems , achieving $O(d)$ computational complexity. The empirical study with FVGC bench mark datasets verifies that our method is both effective and efficient compared to the state-of-the-art FGVC approaches.
****

Saturation-Preserving Specular Reflection Separation
Yuanliu Liu, Zejian Yuan, Nanning Zheng, Yang Wu; Proceedings of the IEEE Confer ence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3725-3733
Specular reflection generally decreases the saturation of surface colors, which will be possibly confused with other colors that have the same hue but lower sat uration. Traditional methods for specular reflection separation suffer this prob lem of hue-saturation ambiguity, producing over-saturated specular-free images q uite often. We proposed a two-step approach to solve this problem. In the first step, we produce an over-saturated specular-free image by global chromaticity pr opagation from specular-free pixels to highlighted ones. Then we recover the sat uration based on priors of the piecewise constancy of diffuse chromaticity as we ll as the spatial sparsity and smoothness of specular reflection. We achieve thi s through increasing the achromatic component of diffuse chromaticity, while the magnitudes of increments are determined by linear programming under the constra ints derived from the priors. Experiments on both laboratory and natural images show that our method can separate the specular reflection while preserving the s aturation of the underlying surface colors.
****

Joint SFM and Detection Cues for Monocular 3D Localization in Road Scenes
Shiyu Song, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3734-3742
We present a system for fast and highly accurate 3D localization of objects like cars in autonomous driving applications, using a single camera. Our localizatio n framework jointly uses information from complementary modalities such as struc ture from motion (SFM) and object detection to achieve high localization accurac y in both near and far fields. This is in contrast to prior works that rely pure ly on detector outputs, or motion segmentation based on sparse feature tracks. R ather than completely commit to tracklets generated by a 2D tracker, we make nov el use of raw detection scores to allow our 3D bounding boxes to adapt to better quality 3D cues. To extract SFM cues, we demonstrate the advantages of dense tr acking over sparse mechanisms in autonomous driving scenarios. In contrast to co mplex scene understanding, our formulation for 3D localization is efficient and can be regarded as an extension of sparse bundle adjustment to incorporate objec t detection cues. Experiments on the KITTI dataset show the efficacy of our cues , as well as the accuracy and robustness of our 3D object localization relative to ground truth and prior works.
****

Fisher Vectors Meet Neural Networks: A Hybrid Classification Architecture
Florent Perronnin, Diane Larlus; Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), 2015, pp. 3743-3752
Fisher Vectors (FV) and Convolutional Neural Networks (CNN) are two image classi fication pipelines with different strengths. While CNNs have shown superior accu racy on a number of classification tasks, FV classifiers are typically less cost ly to train and evaluate. We propose a hybrid architecture that combines their s trengths: the first unsupervised layers rely on the FV while the subsequent full y-connected supervised layers are trained with back-propagation. We show experim entally that this hybrid architecture significantly outperforms standard FV syst ems without incurring the high cost that comes with CNNs. We also derive competi tive mid-level features from our architecture that are readily applicable to oth er class sets and even to new tasks.
****
UniHIST: A Unified Framework for Image Restoration With Marginal Histogram Const raints
Xing Mei, Weiming Dong, Bao-Gang Hu, Siwei Lyu; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3753-3761
Marginal histograms provide valuable information for various computer vision pro blems. However, current image restoration methods do not fully exploit the poten tial of marginal histograms, in particular, their role as ensemble constraints o n the marginal statistics of the restored image. In this paper, we introduce a n ew framework, UniHIST, to incorporate marginal histogram constraints into image restoration. The key idea of UniHIST is to minimize the discrepancy between the marginal histograms of the restored image and the reference histograms in pixel or gradient domains using the quadratic Wasserstein (W2) distance. The W2 distan ce can be computed directly from data without resorting to density estimation. I t provides a differentiable metric between marginal histograms and allows easy i ntegration with existing image restoration methods. We demonstrate the effective ness of UniHIST through denoising of pattern images and non-blind deconvolution of natural images. We show that UniHIST enhances restoration performance and lea ds to visual and quantitative improvements over existing state-of-the-art method s.
****
Human Action Segmentation With Hierarchical Supervoxel Consistency
Jiasen Lu, ran Xu, Jason J. Corso; Proceedings of the IEEE Conference on Compute r Vision and Pattern Recognition (CVPR), 2015, pp. 3762-3771
Detailed analysis of human action, such as action classification, detection and localization has received increasing attention from the community; datasets lik e JHMDB have made it plausible to conduct studies analyzing the impact that such deeper information has on the greater action understanding problem. However, d etailed automatic segmentation of human action has comparatively been unexplored . In this paper, we take a step in that direction and propose a hierarchical MR F model to bridge low-level video fragments with high-level human motion and app earance; novel higher-order potentials connect different levels of the supervoxe l hierarchy to enforce the consistency of the human segmentation by pulling from different segment-scales. Our single layer model significantly outperforms the current state-of-the-art on actionness, and our full model improves upon the s ingle layer baselines in action segmentation.
****
Robust Manhattan Frame Estimation From a Single RGB-D Image
Bernard Ghanem, Ali Thabet, Juan Carlos Niebles, Fabian Caba Heilbron; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 5, pp. 3772-3780
This paper proposes a new framework for estimating the Manhattan Frame (MF) of a n indoor scene from a single RGB-D image. Our technique formulates this problem as the estimation of a rotation matrix that best aligns the normals of the captu red scene to a canonical world axes. By introducing sparsity constraints, our me thod can simultaneously estimate the scene MF, the surfaces in the scene that ar e best aligned to one of three coordinate axes, and the outlier surfaces that d o not align with any of the axes. To test our approach, we contribute a new set of annotations to determine ground truth MFs in each image of the popular NYUv2

dataset. We use this new benchmark to experimentally demonstrate that our metho d is more accurate, faster, more reliable and more robust than the methods used in the literature. We further motivate our technique by showing how it can be us ed to address the RGB-D SLAM problem in indoor scenes by incorporating it into a nd improving the performance of a popular RGB-D SLAM method.
****

Learning to Segment Under Various Forms of Weak Supervision
Jia Xu, Alexander G. Schwing, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3781-3790
Despite the promising performance of conventional fully supervised algorithms, semantic segmentation has remained an important, yet challenging task. Due to t he limited availability of complete annotations, it is of great interest to desi gn solutions for semantic segmentation that take into account weakly labeled dat a, which is readily available at a much larger scale. Contrasting the common the me to develop a different algorithm for each type of weak annotation, in this wo rk, we propose a unified approach that incorporates various forms of weak superv ision -- image level tags, bounding boxes, and partial labels -- to produce a pi xel-wise labeling. We conduct a rigorous evaluation on the challenging Siftflow dataset for various weakly labeled settings, and show that our approach outperfo rms the state-of-the-art by 12% on per-class accuracy, while maintaining com parable per-pixel accuracy.
****

Fast and Accurate Image Upscaling With Super-Resolution Forests
Samuel Schulter, Christian Leistner, Horst Bischof; Proceedings of the IEEE Conf erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3791-3799
The aim of single image super-resolution is to reconstruct a high-resolution ima ge from a single low-resolution input. Although the task is ill-posed it can be seen as finding a non-linear mapping from a low to high-dimensional space. Recen t methods that rely on both neighborhood embedding and sparse-coding have led to tremendous quality improvements. Yet, many of the previous approaches are hard to apply in practice because they are either too slow or demand tedious paramete r tweaks. In this paper, we propose to directly map from low to high-resolutio n patches using random forests. We show the close relation of previous work on s ingle image super-resolution to locally linear regression and demonstrate how ra ndom forests nicely fit into this framework. During training the trees, we optim ize a novel and effective regularized objective that not only operates on the ou tput space but also on the input space, which especially suits the regression ta sk. During inference, our method comprises the same well-known computational eff iciency that has made random forests popular for many computer vision problems.
In the experimental part, we demonstrate on standard benchmarks for single ima ge super-resolution that our approach yields highly accurate state-of-the-art re sults, while being fast in both training and evaluation.
****

Light Field From Micro-Baseline Image Pair
Zhoutong Zhang, Yebin Liu, Qionghai Dai; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2015, pp. 3800-3809
We present a novel phase-based approach for recon- structing 4D light field from a micro-baseline stereo (LfS) pair. Our approach takes advantage of the unique prop- erty of complex steerable pyramid filters in micro-baseline stereo. We fir st introduce a disparity assisted phase based synthesis (DAPS) strategy that can integrate disparity infor- mation into the phase term of a reference image to w arp it to its close neighbor views. Based on the DAPS, an "analy- sis by synthes is" approach is proposed to warp from one of the input binocular images to the o ther, and iteratively op- timize the disparity map to minimize the phase differe nces between the warped one and the ground truth input. Fi- nally, the densely a nd regularly spaced, high quality light field images can be reconstructed using the proposed DAPS according to the refined disparity map. Our approach also solv es the problems of disparity inconsistency and ringing artifact in available pha se-based view synthesis methods. Experimental results demonstrate that our appro ach sub- stantially improves both the quality of disparity map and light field,

compared with the state-of-the-art stereo match- ing and image based rendering a pproaches.
****

Efficient ConvNet-Based Marker-Less Motion Capture in General Scenes With a Low Number of Cameras
Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchuli n, Micha Andriluka, Chris Bregler, Bernt Schiele, Christian Theobalt; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 , pp. 3810-3818
We present a novel method for accurate marker-less capture of articulated skelet on motion of several subjects in general scenes, indoors and outdoors, even from input filmed with as few as two cameras. Our approach unites a discriminative i mage-based joint detection method with a model-based generative motion tracking algorithm through a combined pose optimization energy. The discriminative part-b ased pose detection method, implemented using Convolutional Networks (ConvNet), estimates unary potentials for each joint of a kinematic skeleton model. These u nary potentials are used to probabilistically extract pose constraints for track ing by using weighted sampling from a pose posterior guided by the model. In the final energy, these constraints are combined with an appearance-based model-to- image similarity term. Poses can be computed very efficiently using iterative lo cal optimization, as ConvNet detection is fast, and our formulation yields a com bined pose estimation energy with analytic derivatives. In combination, this ena bles to track full articulated joint angles at state-of-the-art accuracy and tem poral stability with a very low number of cameras.
****

Learning Scene-Specific Pedestrian Detectors Without Real Data
Hironori Hattori, Vishnu Naresh Boddeti, Kris M. Kitani, Takeo Kanade; Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 201 5, pp. 3819-3827
We consider the problem of designing a scene-specific pedestrian detector in a s cenario where we have zero instances of real pedestrian data (i.e., no labeled r eal data or unsupervised real data). This scenario may arise when a new surveill ance system is installed in a novel location and a scene-specific pedestrian det ector must be trained prior to any observations of pedestrians. The key idea of our approach is to infer the potential appearance of pedestrians using geometric scene data and a customizable database of virtual simulations of pedestrian mot ion. We propose an efficient discriminative learning method that generates a spa tially-varying pedestrian appearance model that takes into the account the persp ective geometry of the scene. As a result, our method is able to learn a unique pedestrian classifier customized for every possible location in the scene. Our e xperimental results show that our proposed approach outperforms classical pedest rian detection models and hybrid synthetic-real models. Our results also yield a surprising result, that our method using purely synthetic data is able to outpe rform models trained on real scene-specific data when data is limited.
****

Deep Filter Banks for Texture Recognition and Segmentation
Mircea Cimpoi, Subhransu Maji, Andrea Vedaldi; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3828-3836
Research in texture recognition often concentrates on the problem of material re cognition in uncluttered conditions, an assumption rarely met by applications. I n this work we conduct a first study of material and describable texture attribu tes recognition in clutter, using a new dataset derived from the OpenSurface tex ture repository. Motivated by the challenge posed by this problem, we propose a new texture descriptor, \dcnn, obtained by Fisher Vector pooling of a Convolutio nal Neural Network (CNN) filter bank. \dcnn substantially improves the state-of- the-art in texture, material and scene recognition. Our approach achieves 79.8\% accuracy on Flickr material dataset and 81\% accuracy on MIT indoor scenes, pro viding absolute gains of more than 10\% over existing approaches. \dcnn easily t ransfers across domains without requiring feature adaptation as for methods that build on the fully-connected layers of CNNs. Furthermore, \dcnn can seamlessly

incorporate multi-scale information and describe regions of arbitrary shapes and sizes. Our approach is particularly suited at localizing ``stuff'' categories and obtains state-of-the-art results on MSRC segmentation dataset, as well as promising results on recognizing materials and surface attributes in clutter on the OpenSurfaces dataset.

****

Multiple Random Walkers and Their Application to Image Cosegmentation

Chulwoo Lee, Won-Dong Jang, Jae-Young Sim, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3837-3845

A graph-based system to simulate the movements and interactions of multiple random walkers (MRW) is proposed in this work. In the MRW system, multiple agents traverse a single graph simultaneously. To achieve desired interactions among those agents, a restart rule can be designed, which determines the restart distribution of each agent according to the probability distributions of all agents. In particular, we develop the repulsive rule for data clustering. We illustrate that the MRW clustering can segment real images reliably. Furthermore, we propose a novel image cosegmentation algorithm based on the MRW clustering. Specifically, the proposed algorithm consists of two steps: inter-image concurrence computation and intra-image MRW clustering. Experimental results demonstrate that the proposed algorithm provides promising cosegmentation performance.

****

Beyond the Shortest Path : Unsupervised Domain Adaptation by Sampling Subspaces Along the Spline Flow

Rui Caseiro, Joao F. Henriques, Pedro Martins, Jorge Batista; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3846-3854

Recently, a particular paradigm [9] in the domain adaptation field has received considerable attention by introducing novel and important insights to the problem. In this case, the source/target domains are represented in the form of subspaces, which are treated as points on the Grassmann manifold. The geodesic curve between them is sampled to obtain intermediate points. Then a classifier is learnt using the projections of the data onto these subspaces. Despite its relevance and popularity, this paradigm [9] contains some limitations. Firstly, in real-world applications, that simple curve (i.e. shortest path) does not provide the necessary flexibility to model the domain shift between the training and testing data sets. Secondly, by using the geodesic curve, we are restricted to only one source domain, which does not allow to take fully advantage of the multiple datasets that are available nowadays. It is then, natural to ask whether this popular paradigm could be extended to deal with more complex curves (e.g. splines) and to integrate multi-sources domains. This is a hard problem considering the Riemannian structure of the space, but we propose a mathematically well-founded idea that enables us to solve it. We exploit the geometric insight of rolling maps [14] to compute a spline curve on the Grassmann manifold. The benefits of the proposed idea are demonstrated through several empirical studies on standard datasets. This novel paradigm allows to explicitly integrate multi-source domains while the previous one [9] uses the mean of all sources. This enables to model better the domain shift and take fully advantage of the training datasets.

****

Spherical Embedding of Inlier Silhouette Dissimilarities

Etai Littwin, Hadar Averbuch-Elor, Daniel Cohen-Or; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3855-3863

In this paper, we introduce a spherical embedding technique to position a given set of silhouettes of an object as observed from a set of cameras arbitrarily positioned around the object. Our technique estimates dissimilarities among the silhouettes and embeds them directly in the rotation space SO(3). The embedding is obtained by an optimization scheme applied over the rotations represented with exponential maps. Since the measure for inter-silhouette dissimilarities contains many outliers, our key idea is to perform the embedding by only using a subset of the estimated dissimilarities. We present a technique that carefully scree

ns for inlier-distances, and the pairwise scaled dissimilarities are embedded in a spherical space, diffeomorphic to SO(3). We show that our method outperforms spherical MDS embedding, demonstrate its performance on various multi-view sets, and highlight its robustness to outliers.
****

Semantics-Preserving Hashing for Cross-View Retrieval
Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3864-3872
With benefits of low storage costs and high query speeds, hashing methods are widely researched for efficiently retrieving large-scale data, which commonly contains multiple views, e.g. a news report with images, videos and texts. In this paper, we study the problem of cross-view retrieval and propose an effective Semantics-Preserving Hashing method, termed SePH. Given semantic affinities of training data as supervised information, SePH transforms them into a probability distribution and approximates it with to-be-learnt hash codes in Hamming space via minimizing the Kullback-Leibler divergence. Then kernel logistic regression with a sampling strategy is utilized to learn the nonlinear projections from features in each view to the learnt hash codes. And for any unseen instance, predicted hash codes and their corresponding output probabilities from observed views are utilized to determine its unified hash code, using a novel probabilistic approach. Extensive experiments conducted on three benchmark datasets well demonstrate the effectiveness and reasonableness of SePH.
****

Object Proposal by Multi-Branch Hierarchical Segmentation
Chaoyang Wang, Long Zhao, Shuang Liang, Liqing Zhang, Jinyuan Jia, Yichen Wei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3873-3881
Hierarchical segmentation based object proposal methods have become an important step in modern object detection paradigm. However, standard single-way hierarchical methods are fundamentally flawed in that the errors in early steps cannot be corrected and accumulate. In this work, we propose a novel multi-branch hierarchical segmentation approach that alleviates such problems by learning multiple merging strategies in each step in a complementary manner, such that errors in one merging strategy could be corrected by the others. Our approach achieves the state-of-the-art performance for both object proposal and object detection tasks, comparing to previous object proposal methods.
****

Ambient Occlusion via Compressive Visibility Estimation
Wei Yang, Yu Ji, Haiting Lin, Yang Yang, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3882-3889
There has been emerging interest on recovering traditionally challenging intrinsic scene properties. In this paper, we present a novel computational imaging solution for recovering the ambient occlusion (AO) map of an object. AO measures how much light from all different directions can reach a surface point without being blocked by self-occlusions. Previous approaches either require obtaining highly accurate surface geometry or acquiring a large number of images. We adopt a compressive sensing framework that captures the object under strategically coded lighting directions. We show that this incident illumination field exhibits some unique properties suitable for AO recovery: every ray's contribution to the visibility function is binary while their distribution for AO measurement is sparse. This enables a sparsity-prior based solution for iteratively recovering the surface normal, the surface albedo, and the visibility function from a small number of images. To physically implement the scheme, we construct an encodable directional light source using the light field probe. Experiments on synthetic and real scenes show that our approach is both reliable and accurate with significantly reduced size of input.
****

Shape-Tailored Local Descriptors and Their Application to Segmentation and Tracking

Naeemullah Khan, Marei Algarni, Anthony Yezzi, Ganesh Sundaramoorthi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3890-3899

We propose new dense descriptors for texture segmentation. Given a region of arbitrary shape in an image, these descriptors are formed from shape-dependent scale spaces of oriented gradients. These scale spaces are defined by Poisson-like partial differential equations. A key property of our new descriptors is that they do not aggregate image data across the boundary of the region, in contrast to existing descriptors based on aggregation of oriented gradients. As an example, we show how the descriptor can be incorporated in a Mumford-Shah energy for texture segmentation. We test our method on several challenging datasets for texture segmentation and textured object tracking. Experiments indicate that our descriptors lead to more accurate segmentation than non-shape dependent descriptors and the state-of-the-art in texture segmentation.

****

Scalable Object Detection by Filter Compression With Regularized Sparse Coding

Ting-Hsuan Chao, Yen-Liang Lin, Yin-Hsi Kuo, Winston H. Hsu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3900-3907

For practical applications, an object detection system requires huge number of classes to meet real world needs. Many successful object detection systems use part-based model which trains several filters (classifiers) for each class to perform multiclass object detection. However, these methods have linear computational complexity in regard to the number of classes and may lead to huge computing time. To solve the problem, some works learn a codebook for the filters and conduct operations only on the codebook to make computational complexity sublinear in regard to the number of classes. But the past studies missed to consider filter characteristics, e.g., filters are weights trained by Support Vector Machine, and rather they applied method such as sparse coding for visual signals' optimization. This misuse results in huge accuracy loss when a large speedup is required. To remedy this shortcoming, we have developed a new method called Regularized Sparse Coding which is designed to reconstruct filter functionality. That is, it reconstructs the ability of filter to produce accurate score for classification. Our method can reconstruct filters by minimizing score map error, while sparse coding reconstructs filters by minimizing appearance error. This different optimization strategy makes our method be able to have small accuracy loss when a large speedup is achieved. On the ILSVRC 2013 dataset, which has 200 classes, this work represents a 16 times speedup using only 1.25% memory on single CPU with 0.04 mAP drop when compared with the original Deformable Part Model. Moreover, parallel computing on GPUs is also applicable for our work to achieve more speedup.

****

An Improved Deep Learning Architecture for Person Re-Identification

Ejaz Ahmed, Michael Jones, Tim K. Marks; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3908-3916

In this work we propose a method for simultaneously learning features and a corresponding similarity metric for person re-identification. We present a deep convolution architecture with layers specially designed to address the problem of re-identification. Given a pair of images as input, our network outputs a similarity value indicating whether the two input images depict the same person. Novel elements of our architecture include a layer that computes cross-input neighborhood differences, which capture local relationships among mid-level features that were computed separately from the two input images. A high-level summary of the outputs of this layer is computed by a layer of patch summary features, which are then spatially integrated in subsequent layers. Our method significantly outperforms the state of the art on both a large data set (CUHK03) and a medium-sized dataset (CUHK01), and it is resistant to overfitting. We also demonstrate that by initially training on an unrelated large data set before fine tuning on a small target data set, our network can achieve results comparable to the state of the art even on the small data set (VIPeR).

****
Understanding Classifier Errors by Examining Influential Neighbors
Mayank Kabra, Alice Robie, Kristin Branson; Proceedings of the IEEE Conference o
n Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3917-3925
Modern supervised learning algorithms can learn very accurate and complex discri
minating functions. But when these classifiers fail, this complexity can also be
 a drawback because there is no easy, intuitive way to diagnose why they are fai
ling and remedy the problem. This important question has received little attenti
on. To address this problem, we propose a novel method to analyze and understand
 a classifier's errors. Our method centers around a measure of how much influenc
e a training example has on the classifier's prediction for a test example. To u
nderstand why a classifier is mispredicting the label of a given test example, t
he user can find and review the most influential training examples that caused t
his misprediction, allowing them to focus their attention on relevant areas of t
he data space. This will aid the user in determining if and how the training dat
a is inconsistently labeled or lacking in diversity, or if the feature represent
ation is insufficient. As computing the influence of each training example is co
mputationally impractical, we propose a novel distance metric to approximate inf
luence for boosting classifiers that is fast enough to be used interactively. We
 also show several novel use paradigms of our distance metric. Through experimen
ts, we show that it can be used to find incorrectly or inconsistently labeled tr
aining examples, to find specific areas of the data space that need more trainin
g data, and to gain insight into which features are missing from the current rep
resentation.
****
Riemannian Coding and Dictionary Learning: Kernels to the Rescue
Mehrtash Harandi, Mathieu Salzmann; Proceedings of the IEEE Conference on Comput
er Vision and Pattern Recognition (CVPR), 2015, pp. 3926-3935
While sparse coding on non-flat Riemannian manifolds has recently become increas
ingly popular, existing solutions either are dedicated to specific manifolds, or
 rely on optimization problems that are difficult to solve, especially when it c
omes to dictionary learning. In this paper, we propose to make use of kernels to
 perform coding and dictionary learning on Riemannian manifolds. To this end, we
 introduce a general Riemannian coding framework with its kernel-based counterpa
rt. This lets us (i) generalize beyond the special case of sparse coding; (ii) i
ntroduce efficient solutions to two coding schemes; (iii) learn the kernel param
eters; (iv) perform unsupervised and supervised dictionary learning in a much si
mpler manner than previous Riemannian coding methods. We demonstrate the effecti
veness of our approach on three different types of non-flat manifolds, and illus
trate its generality by applying it to Euclidean spaces, which also are Riemanni
an manifolds.
****
Scalable Structure From Motion for Densely Sampled Videos
Benjamin Resch, Hendrik P. A. Lensch, Oliver Wang, Marc Pollefeys, Alexander Sor
kine-Hornung; Proceedings of the IEEE Conference on Computer Vision and Pattern
Recognition (CVPR), 2015, pp. 3936-3944
Videos consisting of thousands of high resolution frames are challenging for exi
sting structure from motion (SfM) and simultaneous-localization and mapping (SLA
M) techniques. We present a new approach for simultaneously computing extrinsic
 camera poses and 3D scene structure that is capable of handling such large volum
es of image data. The key insight behind this paper is to effectively exploit co
herence in densely sampled video input. Our technical contributions include robu
st tracking and selection of confident video frames, a novel window bundle adjus
tment, frame-to-structure verification for globally consistent reconstructions w
ith multi-loop closing, and utilizing efficient global linear camera pose estima
tion in order to link both consecutive and distant bundle adjustment windows. To
 our knowledge we describe the first system that is capable of handling high res
olution, high frame-rate video data with close to realtime performance. In addit
ion, our approach can robustly integrate data from different video sequences, al
lowing multiple video streams to be simultaneously calibrated in an efficient an

d globally optimal way. We demonstrate high quality alignment on large scale cha
llenging datasets, e.g., 2-20 megapixel resolution at frame rates of 25-120 Hz w
ith thousands of frames.
****

Parsing Occluded People by Flexible Compositions
Xianjie Chen, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2015, pp. 3945-3954
This paper presents an approach to parsing humans when there is significant occl
usion. We model humans using a graphical model which has a tree structure buildi
ng on recent work [32, 6] and exploit the connectivity prior that, even in prese
nce of occlusion, the visible nodes form a connected subtree of the graphical mo
del. We call each connected subtree a flexible composition of object parts. This
 involves a novel method for learning occlusion cues. During inference we need t
o search over a mixture of different flexible models. By exploiting part sharing
, we show that this inference can be done extremely efficiently requiring only t
wice as many computations as searching for the entire object (i.e., not modeling
 occlusion). We evaluate our model on the standard benchmarked "We Are Family" S
tickmen dataset and obtain significant performance improvements over the best al
ternative algorithms.
****

Joint Calibration of Ensemble of Exemplar SVMs
Davide Modolo, Alexander Vezhnevets, Olga Russakovsky, Vittorio Ferrari; Proceed
ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2
015, pp. 3955-3963
We present a method for calibrating the Ensemble of Exemplar SVMs model. Unlike
the standard approach, which calibrates each SVM independently, our method optim
izes their joint performance as an ensemble. We formulate joint calibration as a
 constrained optimization problem and devise an efficient optimization algorithm
 to find its global optimum. The algorithm dynamically discards parts of the sol
ution space that cannot contain the optimum early on, making the optimization co
mputationally feasible. We experiment with EE-SVM trained on state-of-the-art CN
N descriptors. Results on the ILSVRC 2014 and PASCAL VOC 2007 datasets show that
 (i) our joint calibration procedure outperforms independent calibration on the
task of classifying windows as belonging to an object class or not; and (ii) thi
s improved window classifier leads to better performance on the object detection
 task.
****

Holistic 3D Scene Understanding From a Single Geo-Tagged Image
Shenlong Wang, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3964-3972
In this paper we are interested in exploiting geographic priors to help outdoor
scene understanding. Towards this goal we propose a holistic approach that reas
ons jointly about 3D object detection, pose estimation, semantic segmentation as
 well as depth reconstruction from a single image. Our approach takes advantage
 of large-scale crowd-sourced maps to generate dense geographic, geometric and
semantic priors by rendering the 3D world. We demonstrate the effectiveness of o
ur holistic model on the challenging KITTI dataset, and show significant improve
ments over the baselines in all metrics and tasks.
****

A Large-Scale Car Dataset for Fine-Grained Categorization and Verification
Linjie Yang, Ping Luo, Chen Change Loy, Xiaoou Tang; Proceedings of the IEEE Con
ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3973-3981
This paper aims to highlight vision related tasks centered around "car", which h
as been largely neglected by vision community in comparison to other objects. We
 show that there are still many interesting car-related problems and application
s, which are not yet well explored and researched. To facilitate future car-rela
ted research, in this paper we present our on-going effort in collecting a large
-scale dataset, "CompCars", that covers not only different car views, but also t
heir different internal and external parts, and rich attributes. Importantly, th
e dataset is constructed with a cross-modality nature, containing a surveillance

nature set and a web-nature set. We further demonstrate a few important applicat
ions exploiting the dataset, namely car model classification, car model verifica
tion, and attribute prediction. We also discuss specific challenges of the car-r
elated problems and other potential applications that worth further investigatio
ns. The latest dataset can be downloaded at http://mmlab.ie.cuhk.edu.hk/datasets
/comp_cars/index.html
****

DeepContour: A Deep Convolutional Feature Learned by Positive-Sharing Loss for C
ontour Detection
Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, Zhijiang Zhang; Proceedings of the
 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 39
82-3991
Contour detection serves as the basis of a variety of computer vision tasks such
 as image segmentation and object recognition. The mainstream works to address t
his problem focus on designing engineered gradient features. In this work, we sh
ow that contour detection accuracy can be improved by instead making the use of
the deep features learned from convolutional neural networks (CNNs). While rathe
r than using the networks as a blackbox feature extractor, we customize the trai
ning strategy by partitioning contour (positive) data into subclasses and fittin
g each subclass by different model parameters. A new loss function, named positi
ve-sharing loss, in which each subclass shares the loss for the whole positive c
lass, is proposed to learn the parameters. Compared to the sofmax loss function,
 the proposed one, introduces an extra regularizer to emphasizes the losses for
the positive and negative classes, which facilitates to explore more discriminat
ive features. Our experimental results demonstrate that learned deep features ca
n achieve top performance on Berkeley Segmentation Dataset and Benchmark (BSDS50
0) and obtain competitive cross dataset generalization result on the NYUD datase
t.
****

Convolutional Feature Masking for Joint Object and Stuff Segmentation
Jifeng Dai, Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2015, pp. 3992-4000
The topic of semantic segmentation has witnessed considerable progress due to th
e powerful features learned by convolutional neural networks (CNNs). The current
 leading approaches for semantic segmentation exploit shape information by extra
cting CNN features from masked image regions. This strategy introduces artificia
l boundaries on the images and may impact the quality of the extracted features.
 Besides, the operations on the raw image domain require to compute thousands of
 networks on a single image, which is time-consuming. In this paper, we propose
 to exploit shape information via masking convolutional features. The proposal s
egments (e.g., super-pixels) are treated as masks on the convolutional feature m
aps. The CNN features of segments are directly masked out from these maps and us
ed to train classifiers for recognition. We further propose a joint method to ha
ndle objects and "stuff" (e.g., grass, sky, water) in the same framework. State-
of-the-art results are demonstrated on the challenging PASCAL VOC benchmarks, wi
th a compelling computational speed.
****

A Fixed Viewpoint Approach for Dense Reconstruction of Transparent Objects
Kai Han, Kwan-Yee K. Wong, Miaomiao Liu; Proceedings of the IEEE Conference on C
omputer Vision and Pattern Recognition (CVPR), 2015, pp. 4001-4008
This paper addresses the problem of reconstructing the surface shape of transpar
ent objects. The difficulty of this problem originates from the viewpoint depend
ent appearance of a transparent object, which quickly makes reconstruction metho
ds tailored for diffuse surfaces fail disgracefully. In this paper, we develop a
 fixed viewpoint approach for dense surface reconstruction of transparent object
s based on refraction of light. We introduce a simple setup that allows us alter
 the incident light paths before light rays enter the object, and develop a meth
od for recovering the object surface based on reconstructing and triangulating s
uch incident light paths. Our proposed approach does not need to model the compl
ex interactions of light as it travels through the object, neither does it assum

e any parametric form for the shape of the object nor the exact number of refrac
tions and reflections taken place along the light paths. It can therefore handle
 transparent objects with a complex shape and structure, with unknown and even i
nhomogeneous refractive index. Experimental results on both synthetic and real d
ata are presented which demonstrate the feasibility and accuracy of our proposed
 approach.
****

Low-Level Vision by Consensus in a Spatial Hierarchy of Regions
Ayan Chakrabarti, Ying Xiong, Steven J. Gortler, Todd Zickler; Proceedings of th
e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4
009-4017
We introduce a multi-scale framework for low-level vision, where the goal is est
imating physical scene values from image data---such as depth from stereo image
pairs. The framework uses a dense, overlapping set of image regions at multiple
scales and a ``local model,'' such as a slanted-plane model for stereo disparity
, that is expected to be valid piecewise across the visual field. Estimation is
cast as optimization over a dichotomous mixture of variables, simultaneously det
ermining which regions are inliers with respect to the local model (binary varia
bles) and the correct co-ordinates in the local model space for each inlying reg
ion (continuous variables). When the regions are organized into a multi-scale hi
erarchy, optimization can occur in an efficient and parallel architecture, where
 distributed computational units iteratively perform calculations and share info
rmation through sparse connections between parents and children. The framework
performs well on a standard benchmark for binocular stereo, and it produces a di
stributional scene representation that is appropriate for combining with higher-
level reasoning and other low-level cues.
****

Line Drawing Interpretation in a Multi-View Context
Jean-Dominique Favreau, Florent Lafarge, Adrien Bousseau; Proceedings of the IEE
E Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4018-4
026
Many design tasks involve the creation of new objects in the context of an exist
ing scene. Existing work in computer vision only provides partial support for su
ch tasks. On the one hand, multi-view stereo algorithms allow the reconstruction
 of real-world scenes, while on the other hand algorithms for line-drawing inter
pretation do not take context into account. Our work combines the strength of th
ese two domains to interpret line drawings of imaginary objects drawn over photo
graphs of an existing scene. The main challenge we face is to identify the exist
ing 3D structure that correlates with the line drawing while also allowing the c
reation of new structure that is not present in the real world. We propose a lab
eling algorithm to tackle this problem, where some of the labels capture dominan
t orientations of the real scene while a free label allows the discovery of new
orientations in the imaginary scene. We illustrate our algorithm by interpreting
 line drawings for urban planing, home remodeling, furniture design and cultural
 heritage.
****

Toward User-Specific Tracking by Detection of Human Shapes in Multi-Cameras
Chun-Hao Huang, Edmond Boyer, Bibiana do Canto Angonese, Nassir Navab, Slobodan
Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2015, pp. 4027-4035
Human shape tracking consists in fitting a template model to temporal sequences
of visual observations. It usually comprises an association step, that finds cor
respondences between the model and the input data, and a deformation step, that
fits the model to the observations given correspondences. Most current approache
s find their common ground with the Iterative-Closest-Point (ICP) algorithm, whi
ch facilitates the association step with local distance considerations. It fails
 when large deformations occur, and errors in the association tend to propagate
over time. In this paper, we propose a discriminative alternative for the associ
ation, that leverages random forests to infer correspondences in one shot. It al
lows for large deformations and prevents tracking errors from accumulating. The

approach is successfully integrated to a surface tracking framework that recover s human shapes and poses jointly. When combined with ICP, this discriminative as sociation proves to yield better accuracy in registration, more stability when t racking over time, and faster convergence. Evaluations on existing datasets demo nstrate the benefits with respect to the state-of-the-art.
****

Intra-Frame Deblurring by Leveraging Inter-Frame Camera Motion
Haichao Zhang, Jianchao Yang; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2015, pp. 4036-4044
Camera motion introduces motion blur, degrading the quality of video. A video d eblurring method is proposed based on two observations: (i) camera motion within capture of each individual frame leads to motion blur; (ii) camera motion betwe en frames yields inter-frame mis-alignment that can be exploited for blur remo val. The proposed method effectively leverages the information distributed acros s multiple video frames due to camera motion, jointly estimating the motion bet ween consecutive frames and blur within each frame. This joint analysis is cruci al for achieving effective restoration by leveraging temporal information. Exten sive experiments are carried out on synthetic data as well as real-world blurry videos. Comparisons with several state-of-the-art methods verify the effectiven ess of the proposed method.
****

Salient Object Subitizing
Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe L in, Xiaohui Shen, Brian Price, Radomir Mech; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4045-4054
People can immediately and precisely identify 1, 2, 3 or 4 items by a simple gla nce. The phenomenon, known as Subitizing, inspires us to pursue the task of Sali ent Object Subitizing (SOS), i.e. predicting the existence and the number of sal ient objects in a scene using holistic cues. To study this problem, we propose a new image dataset annotated by Amazon Mechanical Turk. We show that for a sub stantial proportion of our dataset, there is a high labeling consistency among different subjects, even when a very limited viewing time (0.5s) is given. On ou r dataset, the baseline method using the global Convolutional Neural Network (CN N) feature achieves 94% recall rate in detecting the existence of salient object s, and 42-82% recall rate (chance is 20%) in predicting the number of salient o bjects (1, 2, 3, and 4+), without resorting to any object localization process. Finally, we demonstrate the usefulness of the proposed subitizing technique in t wo computer vision applications: salient object detection and object proposal.
****

Hierarchical-PEP Model for Real-World Face Recognition
Haoxiang Li, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4055-4064
Pose variation remains one of the major factors adversely affect the accuracy of real-world face recognition systems. Inspired by the recently proposed probabil istic elastic part (PEP) model and the success of the deep hierarchical architec ture in a number of visual tasks, we propose the Hierarchical-PEP model to appro ach the unconstrained face recognition problem. We apply the PEP model hierarchi cally to decompose a face image into face parts at different levels of details t o build pose-invariant part-based face representations. Following the hierarchy from bottom-up, we stack the face part representations at each layer, discrimina tively reduce its dimensionality, and hence aggregate the face part representati ons layer-by-layer to build a compact and invariant face representation. The Hie rarchical-PEP model exploits the fine-grained structures of the face parts at di fferent levels of details to address the pose variations. It is also guided by s upervised information in constructing the face part/face representations. We emp irically verify the Hierarchical-PEP model on two public benchmarks (i.e., the L FW and YouTube Faces) and a face recognition challenge (i.e., the PaSC grand cha llenge) for image-based and video-based face verification. The state-of-the-art performance demonstrates the potential of our method.
****

The Common Self-Polar Triangle of Concentric Circles and Its Application to Camera Calibration
Haifei Huang, Hui Zhang, Yiu-ming Cheung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4065-4072
In projective geometry, the common self-polar triangle has often been used to discuss the position relationship of two planar conics. However, there are few researches on the properties of the common self-polar triangle, especially when the two planar conics are special conics. In this paper, we explore the properties of the common self-polar triangle, when the two conics happen to be concentric circles. We show there exist infinite many common self-polar triangles of two concentric circles, and provide a method to locate the vertices of these triangles. By investigating all these triangles, we find that they encode two important properties. The first one is all triangles share one common vertex, and the opposite side of the common vertex lies on the same line, which are the circle center and the line at the infinity of the support plane. The second is all triangles are right triangles. Based on these two properties, the imaged circle center and the varnishing line of support plane can be recovered simultaneously, and many conjugate pairs on vanishing line can be obtained. These allow to induce good constraints on the image of absolute conic. We evaluate two calibration algorithms, whereby accurate results are achieved. The main contribution of this paper is that we initiate a new perspective to look into circle-based camera calibration problem. We believe that other calibration methods using different circle patterns can benefit from this perspective, especially for the patterns which involve more than two circles.
****

Taking a Deeper Look at Pedestrians
Jan Hosang, Mohamed Omran, Rodrigo Benenson, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4073-4082
The only goal of the abstract is to the answer the question: why should I read this paper? In this paper we study the use of convolutional neural networks (convnets) for the task of pedestrian detection. Despite their recent diverse successes, convnets historically underperform compared to other pedestrian detectors. We deliberately omit explicitly modelling the problem into the network (e.g. parts or occlusion modelling) and show that we can reach competitive performance without bells and whistles. In a wide range of experiments we analyse small and big convnets, their architectural choices, parameters, and the influence of different training data, including pre-training on surrogate tasks. We present the best convnet detectors on the Caltech and KITTI dataset. On Caltech our convnets reach top performance both for the Caltech1x and Caltech10x training setup. Using additional data at training time our strongest convnet model is competitive to detectors that use instead additional data at test time.
****

Learning to Segment Moving Objects in Videos
Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4083-4090
We segment moving objects in videos by ranking spatio-temporal segment proposals according to ``moving objectness''; how likely they are to contain a moving object. In each video frame, we compute segment proposals using multiple figure-ground segmentations on per frame motion boundaries. We rank them with a Moving Objectness Detector trained on image and motion fields to detect moving objects and discard over/under segmentations or background parts of the scene. We extend the top ranked segments into spatio-temporal tubes using random walkers on motion affinities of dense point trajectories. Our final tube ranking consistently outperforms previous segmentation methods in the two largest video segmentation benchmarks currently available, for any number of proposals. Further, our per frame moving object proposals increase the detection rate up to 7\% over previous state-of-the-art static proposal methods.
****

GMMCP Tracker: Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking
Afshin Dehghan, Shayan Modiri Assari, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4091-4099
Data association is the backbone to many multiple object tracking (MOT) methods. In this paper we formulate data association as a Generalized Maximum Multi Clique problem (GMMCP). We show that this is the ideal case of modeling tracking in real world scenario where all the pairwise relationships between targets in a batch of frames are taken into account. Previous works assume simplified version of our tracker either in problem formulation or problem optimization. However, we propose a solution using GMMCP where no simplification is assumed in either steps. We show that the NP hard problem of GMMCP can be formulated through Binary-Integer Program where for small and medium size MOT problems the solution can be found efficiently. We further propose a speed-up method, employing Aggregated Dummy Nodes for modeling occlusion and miss-detection, which reduces the size of the input graph without using any heuristics. We show that, using the speedup method, our tracker lends itself to real-time implementation which is plausible in many applications. We evaluated our tracker on six challenging sequences of Town Center, TUD-Crossing, TUD-Stadtmitte, Parking-lot 1, Parking-lot 2 and Parking-lot pizza and show favorable improvement against state of art.
****

Learning Graph Structure for Multi-Label Image Classification via Clique Generation
Mingkui Tan, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Junbin Gao, Fuyuan Hu, Zhen Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4100-4109
Exploiting label dependency for multi-label image classification can significantly improve classification performance. Probabilistic Graphical Models are one of the primary methods for representing such dependencies. The structure of graphical models, however, is either determined heuristically or learned from very limited information. Moreover, neither of these approaches scales well to large or complex graphs. We propose a principled way to learn the structure of a graphical model by considering input features and labels, together with loss functions. We formulate this problem into a max-margin framework initially, and then transform it into a convex programming problem. Finally, we propose a highly scalable procedure that activates a set of cliques iteratively. Our approach exhibits both strong theoretical properties and a significant performance improvement over state-of-the-art methods on both synthetic and real-world data sets.
****

Matrix Completion for Resolving Label Ambiguity
Ching-Hui Chen, Vishal M. Patel, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4110-4118
In real applications, data is not always explicitly-labeled. For instance, label ambiguity exists when we associate two persons appearing in a news photo with two names provided in the caption. We propose a matrix completion-based method for predicting the actual labels from the ambiguously labeled instances, and a standard supervised classifier can learn from the disambiguated labels to classify new data. We further generalize the method to handle the labeling constraints between instances when such prior knowledge is available. Compared to existing methods, our approach achieves 2.9% improvement on the labeling accuracy of the Lost dataset and comparable performance on the Labeled Yahoo! News dataset.
****

Video Magnification in Presence of Large Motions
Mohamed Elgharib, Mohamed Hefeeda, Fredo Durand, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4119-4127
Video magnification reveals subtle variations that would be otherwise invisible to the naked eye. Current techniques require all motion in the video to be very small, which is unfortunately not always the case. Tiny yet meaningful motions are often combined with larger motions, such as the small vibrations of a gate as

it rotates, or the microsaccades in a moving eye. We present a layer-based video magnification approach that can amplify small motions within large ones. An examined region/layer is temporally aligned and subtle variations are magnified. Matting is used to magnify only region of interest while maintaining integrity of nearby sites. Results show handling larger motions, larger amplification factors and significant reduction in artifacts over state of the art.
****

Flying Objects Detection From a Single Moving Camera
Artem Rozantsev, Vincent Lepetit, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4128-4136
We propose an approach to detect flying objects such as UAVs and aircrafts when they occupy a small portion of the field of view, possibly moving against complex backgrounds, and are filmed by a camera that itself moves. Solving such a difficult problem requires combining both appearance and motion cues. To this end we propose a regression-based approach to motion stabilization of local image patches that allows us to achieve effective classifica- tion on spatio-temporal image cubes and outperform state- of-the-art techniques. As the problem is relatively new, we collected two chal- lenging datasets for UAVs and Aircrafts, which can be used as benchmarks for flying objects detection and vision- guided collision avoidance.
****

Line-Based Multi-Label Energy Optimization for Fisheye Image Rectification and Calibration
Mi Zhang, Jian Yao, Menghan Xia, Kai Li, Yi Zhang, Yaping Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4137-4145
Fisheye image rectification and estimation of intrinsic parameters for real scenes have been addressed in the literature by using line information on the distorted images. In this paper, we propose an easily implemented fisheye image rectification algorithm with line constrains in the undistorted perspective image plane. A novel Multi-Label Energy Optimization (MLEO) method is adopted to merge short circular arcs sharing the same or the approximately same circular parameters and select long circular arcs for camera rectification. Further we propose an efficient method to estimate intrinsic parameters of the fisheye camera by automatically selecting three properly arranged long circular arcs from previously obtained circular arcs in the calibration procedure. Experimental results on a number of real images and simulated data show that the proposed method can achieve good results and outperforms the existing approaches and the commercial software in most cases.
****

Adaptive Eye-Camera Calibration for Head-Worn Devices
David Perra, Rohit Kumar Gupta, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4146-4155
We present a novel, continuous, locally optimal calibration scheme for use with head-worn devices. Current calibration schemes solve for a globally optimal model of the eye-device transformation by performing calibration on a per-user or once-per-use basis. However, these calibration schemes are impractical for real-world applications because they do not account for changes in calibration during the time of use. Our calibration scheme allows a head-worn device to calculate a locally optimal eye-device transformation on demand by computing an optimal model from a local window of previous frames. By leveraging naturally occurring interest regions within the user's environment, our system can calibrate itself without the user's active participation. Experimental results demonstrate that our proposed calibration scheme outperforms the existing state of the art systems while being significantly less restrictive to the user and the environment.
****

Modeling Object Appearance Using Context-Conditioned Component Analysis
Daniyar Turmukhambetov, Neill D.F. Campbell, Simon J.D. Prince, Jan Kautz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4156-4164

Subspace models have been very successful at modeling the appearance of structured image datasets when the visual objects have been aligned in the images (e.g., faces). Even with extensions that allow for global transformations or dense warps of the image, the set of visual objects whose appearance may be modeled by such methods is limited. They are unable to account for visual objects where occlusion leads to changing visibility of different object parts (without a strict layered structure) and where a one-to-one mapping between parts is not preserved. For example bunches of bananas contain different numbers of bananas but each individual banana shares an appearance subspace. In this work we remove the image space alignment limitations of existing subspace models by conditioning the models on a shape dependent context that allows for the complex, non-linear structure of the appearance of the visual object to be captured and shared. This allows us to exploit the advantages of subspace appearance models with non-rigid, deformable objects whilst also dealing with complex occlusions and varying numbers of parts. We demonstrate the effectiveness of our new model with examples of structured inpainting and appearance transfer.
****

Displets: Resolving Stereo Ambiguities Using Object Knowledge
Fatma Guney, Andreas Geiger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4165-4175
Stereo techniques have witnessed tremendous progress over the last decades, yet some aspects of the problem still remain challenging today. Striking examples are reflecting and textureless surfaces which cannot easily be recovered using traditional local regularizers. In this paper, we therefore propose to regularize over larger distances using object-category specific disparity proposals (displets) which we sample using inverse graphics techniques based on a sparse disparity estimate and a semantic segmentation of the image. The proposed displets encode the fact that objects of certain categories are not arbitrarily shaped but typically exhibit regular structures. We integrate them as non-local regularizer for the challenging object class 'car' into a superpixel based CRF framework and demonstrate its benefits on the KITTI stereo evaluation. At time of submission, our approach ranks first across all KITTI stereo leaderboards.
****

Time-to-Contact From Image Intensity
Yukitoshi Watanabe, Fumihiko Sakaue, Jun Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4176-4183
It is known that time-to-contact toward objects can be estimated just from changes in the object size in camera images, and we do not need any additional information, such as camera parameters and motions. However, the existing methods for measuring the time-to-contact are based on geometric image features, such as corners and edge lines, and thus they cannot be used when there are no geometric features in images. In this paper, we propose a new method for computing the time-to-contact from photometric information in images. When a light source moves in the 3D scene, an observed intensity changes according to the motion of the light source. In this paper, we analyze the change in photometric information in images, and show that the time-to-contact can be estimated just from the changes in intensity in images. Our method does not need any additional information, such as radiance of light source, reflectance of object and orientation of object surface. The proposed method can be used in various applications, such as vehicle driver assistance.
****

Transferring a Semantic Representation for Person Re-Identification and Search
Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4184-4193
Learning semantic attributes for person re-identification and description-based person search has gained increasing interest due to attributes' great potential as a pose and view-invariant representation. However, existing attribute-centric approaches have thus far underperformed state-of-the-art conventional approaches. This is due to their non-scalable need for extensive domain (camera) specific annotation. In this paper we present a new semantic attribute learning approa

ch for person re-identification and search. Our model is trained on existing fas hion photography datasets -- either weakly or strongly labelled. It can then be transferred and adapted to provide a powerful semantic description of surveillan ce person detections, without requiring any surveillance domain supervision. The resulting representation is useful for both unsupervised and supervised person re-identification, achieving state-of-the-art and near state-of-the-art performa nce respectively. Furthermore, as a semantic representation it allows descriptio n-based person search to be integrated within the same framework.
****

Robust Video Segment Proposals With Painless Occlusion Handling
Zhengyang Wu, Fuxin Li, Rahul Sukthankar, James M. Rehg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4194-42 03
We propose a robust algorithm to generate video segment proposals. The proposals generated by our method can start from any frame in the video and are robust to complete occlusions. Our method does not assume specific motion models and even has a limited capability to generalize across videos. We build on our previous least squares tracking framework, where image segment proposals are generated an d tracked using learned appearance models. The innovation in our new method lies in the use of two efficient moves, the merge move and free addition, to efficie ntly start segments from any frame and track them through complete occlusions, w ithout much additional computation. Segment size interpolation is used for effec tively detecting occlusions. We propose a new metric for evaluating video segme nt proposals on the challenging VSB-100 benchmark and present state-of-the-art r esults. Preliminary results are also shown for the potential use of our framewor k to track segments across different videos.
****

Face Alignment Using Cascade Gaussian Process Regression Trees
Donghoon Lee, Hyunsin Park, Chang D. Yoo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4204-4212
In this paper, we propose a face alignment method that uses cascade Gaussian pro cess regression trees (cGPRT) constructed by combining Gaussian process regressi on trees (GPRT) in a cascade stage-wise manner. Here, GPRT is a Gaussian process with a kernel defined by a set of trees. The kernel measures the similarity bet ween two inputs as the number of trees where the two inputs fall in the same lea ves. Without increasing prediction time, the prediction of cGPRT can be performe d in the same framework as the cascade regression trees (CRT) but with better ge neralization. Features for GPRT are designed using shape-indexed difference of G aussian (DoG) filter responses sampled from local retinal patterns to increase s tability and to attain robustness against geometric variances. Compared with the previous CRT-based face alignment methods that have shown state-of-the-art perf ormances, cGPRT using shape-indexed DoG features performed best on the HELEN and 300-W datasets which are the most challenging dataset today.
****

Regularizing Max-Margin Exemplars by Reconstruction and Generative Models
Jose C. Rubio, Bjorn Ommer; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2015, pp. 4213-4221
Part-based models are one of the leading paradigms in visual recognition. In the absence of costly part annotations, associating and aligning different training instances of a part classifier and finding characteristic negatives is challeng ing and computationally demanding. To avoid this costly mining of training sampl es, we estimate separate generative models for negatives and positives and integ rate them into a max-margin exemplar-based model. The generative model and a spa rsity constraint on the correlation between spatially neighboring feature dimens ions regularize the part filters during learning and improve their generalizatio n to similar instances. To suppress inappropriate positive part samples, we proj ect the classifier back into the image domain and penalize against deviations fr om the original exemplar image patch. The part filter is then optimized to i) di scriminate against clutter, to ii) generalize to similar instances of the part, and iii) to yield a good reconstruction of the original image patch. Moreover, w

e propose an approximation for estimating the geometric margin so that learning large numbers of parts becomes feasible. Experiments show improved part localization, object recognition, and part-based reconstruction performance compared to popular exemplar-based approaches on PASCAL VOC
****

A Fast Algorithm for Elastic Shape Distances Between Closed Planar Curves
Gunay Dogan, Javier Bernal, Charles R. Hagwood; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4222-4230
Effective computational tools for shape analysis are needed in many areas of science and engineering. We address this and propose a new fast iterative algorithm to compute the elastic geodesic distance between shapes of closed planar curves. The original algorithm for this has cubic time complexity with respect to the number of nodes per curve. Hence it is not suitable for large shape data sets. We aim for large-scale shape analysis and thus propose an iterative algorithm based on the original one but with quadratic time complexity. In practice, we observe subquadratic, almost linear running times, and that our algorithm scales very well with large numbers of nodes. The key to our algorithm is the decoupling of the optimization for the starting point and rotation from that of the reparametrization, and the development of fast dynamic programming and iterative nonlinear constrained optimization algorithms that work in tandem to compute optimal reparametrizations fast.
****

Reflection Removal for In-Vehicle Black Box Videos
Christian Simon, In Kyu Park; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4231-4239
In-vehicle black box camera becomes an popular equipment in many countries for security monitoring and event capturing. The readability of video contents is the most important capability, which is, however, often degraded due to the reflection on the windscreen. In this paper, we propose a novel method to remove the reflection on the windscreen in the in-vehicle black box videos. The main idea is to exploit spatio-temporal coherence of the reflection which shows that a vehicle moves forward while the reflection layer of internal object remains static. The average image prior is introduced by imposing a heavy-tail distribution with higher peak. The two layered scene is the base of the separation model. In order to remove the reflection, a cost non-convex function is developed based on this property and optimized. Experimental results demonstrate that the proposed approach successfully separates the layers in real black box videos.
****

Tree Quantization for Large-Scale Similarity Search and Classification
Artem Babenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4240-4248
We propose a new vector encoding scheme (tree quantization) that obtains lossy compact codes for high-dimensional vectors via tree-based dynamic programming. Similarly to several previous schemes such as product quantization, these codes correspond to codeword numbers within multiple codebooks. We propose an integer programming-based optimization that jointly recovers the coding tree structure and the codebooks by minimizing the compression error on a training dataset. In the experiments with diverse visual descriptors (SIFT, neural codes, Fisher vectors), tree quantization is shown to combine fast encoding and state-of-the-art accuracy in terms of the compression error, the retrieval performance, and the image classification error.
****

Integrating Parametric and Non-Parametric Models For Scene Labeling
Bing Shuai, Gang Wang, Zhen Zuo, Bing Wang, Lifan Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4249-4258
We adopt Convolutional Neural Networks (CNN) as our parametric model to learn discriminative features and classifiers for local patch classification. As visually similar pixels are indistinguishable from local context, we alleviate such ambiguity by putting a global scene constraint. We estimate the global potential in

a non-parametric framework. Furthermore, a large margin based CNN metric learni
ng method is proposed for better global potential estimation. The final pixel cl
ass prediction is performed by integrating local and global beliefs. Even withou
t any post-processing, we achieve state-of-the-art on SiftFlow and competitive r
esults on Stanford Background benchmark.
****

Mining Semantic Affordances of Visual Object Categories
Yu-Wei Chao, Zhan Wang, Rada Mihalcea, Jia Deng; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4259-4267
Affordances are fundamental attributes of objects. Affordances reveal the functi
onalities of objects and the possible actions that can be performed on them. Und
erstanding affordances is crucial for recognizing human activities in visual dat
a and for robots to interact with the world. In this paper we introduce the new
 problem of mining the knowledge of semantic affordance: given an object, determ
ining whether an action can be performed on it. This is equivalent to connecting
 verb nodes and noun nodes in WordNet, or filling an affordance matrix encoding
the plausibility of each action-object pair. We introduce a new benchmark with c
rowdsourced ground truth affordances on 20 PASCAL VOC object classes and 957 act
ion classes. We explore a number of approaches including text mining, visual mi
ning, and collaborative filtering. Our analyses yield a number of significant in
sights that reveal the most effective ways of collecting knowledge of semantic a
ffordances.
****

Causal Video Object Segmentation From Persistence of Occlusions
Brian Taylor, Vasiliy Karasev, Stefano Soatto; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4268-4276
Occlusion relations inform the partition of the image domain into ``objects'' bu
t are difficult to determine from a single image or short-baseline video. We sho
w how long-term occlusion relations can be robustly inferred from video, and use
d within a convex optimization framework to segment the image domain into region
s. We highlight the challenges in determining these occluder/occluded relations
and ensuring regions remain temporally consistent, propose strategies to overcom
e them, and introduce an efficient numerical scheme to perform the partition dir
ectly on the pixel grid, without the need for superpixelization or other preproc
essing steps.
****

Multiple Instance Learning for Soft Bags via Top Instances
Weixin Li, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2015, pp. 4277-4285
A generalized formulation of the multiple instance learning problem is considere
d. Under this formulation, both positive and negative bags are soft, in the sens
e that negative bags can also contain positive instances. This reflects a proble
m setting commonly found in practical applications, where labeling noise appears
 on both positive and negative training samples. A novel bag-level representatio
n is introduced, using instances that are most likely to be positive (denoted t
op instances), and its ability to separate soft bags, depending on their relativ
e composition in terms of positive and negative instances, is studied. This stud
y inspires a new large-margin algorithm for soft-bag classification, based on a
latent support vector machine that efficiently explores the combinatorial space
of bag compositions. Empirical evaluation on three datasets is shown to confirm
the main findings of the theoretical analysis and the effectiveness of the prop
osed soft-bag classifier.
****

Multiclass Semantic Video Segmentation With Object-Level Active Inference
Buyu Liu, Xuming He; Proceedings of the IEEE Conference on Computer Vision and P
attern Recognition (CVPR), 2015, pp. 4286-4294
We address the problem of integrating object reasoning with supervoxel labeling
in multiclass semantic video segmentation. To this end, we first propose an obje
ct-augmented dense CRF in spatio-temporal domain, which captures long-range depe
ndency between supervoxels, and imposes consistency between object and supervoxe

l labels. We develop an efficient mean field inference algorithm to jointly infe
r the supervoxel labels, object activations and their occlusion relations for a
moderate number of object proposals. To scale up our method, we adopt an acti
ve inference strategy to improve the efficiency, which adaptively selects object
 subgraphs in the object-augmented dense CRF. We formulate the problem as a Mark
ov Decision Process, which learns an approximate optimal policy based on a rewar
d of accuracy improvement and a set of well-designed model and input features. W
e evaluate our method on three publicly available multiclass video semantic segm
entation datasets and demonstrate superior efficiency and accuracy.
****

Effective Face Frontalization in Unconstrained Images
Tal Hassner, Shai Harel, Eran Paz, Roee Enbar; Proceedings of the IEEE Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4295-4304
"Frontalization" is the process of synthesizing frontal facing views of faces ap
pearing in single unconstrained photos. Recent reports have suggested that this
process may substantially boost the performance of face recognition systems. Thi
s, by transforming the challenging problem of recognizing faces viewed from unco
nstrained viewpoints to the easier problem of recognizing faces in constrained,
forward facing poses. Previous frontalization methods did this by attempting to
approximate 3D facial shapes for each query image. We observe that 3D face shape
 estimation from unconstrained photos may be a harder problem than frontalizatio
n and can potentially introduce facial misalignments. Instead, we explore the si
mpler approach of using a single, unmodified, 3D surface as an approximation to
the shape of all input faces. We show that this leads to a straightforward, effi
cient and easy to implement method for frontalization. More importantly, it prod
uces aesthetic new frontal views and is surprisingly effective when used for fac
e recognition and gender estimation.
****

Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors
Limin Wang, Yu Qiao, Xiaoou Tang; Proceedings of the IEEE Conference on Computer
 Vision and Pattern Recognition (CVPR), 2015, pp. 4305-4314
Visual features are of vital importance for human action understanding in videos
. This paper presents a new video representation, called trajectory-pooled deep-
convolutional descriptor (TDD), which shares the merits of both hand-crafted fea
tures and deep-learned features. Specifically, we utilize deep architectures to
learn discriminative convolutional feature maps, and conduct trajectory-constrai
ned pooling to aggregate these convolutional features into effective descriptors
. To enhance the robustness of TDDs, we design two normalization methods to tran
sform convolutional feature maps, namely spatiotemporal normalization and channe
l normalization. The advantages of our features come from (i) TDDs are automatic
ally learned and contain high discriminative capacity compared with those hand-c
rafted features; (ii) TDDs take account of the intrinsic characteristics of temp
oral dimension and introduce the strategies of trajectory-constrained sampling a
nd pooling for aggregating deep-learned features. We conduct experiments on two
challenging datasets: HMDB51 and UCF101. Experimental results show that TDDs out
perform previous hand-crafted features and deep-learned features. Our method als
o achieves superior performance to the state of the art on these datasets.
****

Weakly Supervised Localization of Novel Objects Using Appearance Transfer
Mrigank Rochan, Yang Wang; Proceedings of the IEEE Conference on Computer Vision
 and Pattern Recognition (CVPR), 2015, pp. 4315-4324
We consider the problem of localizing unseen objects in weakly labeled image col
lections. Given a set of images annotated at the image level, our goal is to loc
alize the object in each image. The novelty of our proposed work is in addition
to building object appearance model from the weakly labeled data, we also make u
se of existing detectors for some other object classes (which we call ``familiar
 objects''). We propose a method for transferring the appearance models of the f
amiliar objects to the unseen object. Our experimental results on both image and
 video datasets demonstrate the effectiveness of our approach.
****

First-Person Pose Recognition Using Egocentric Workspaces
Gregory Rogez, James S. Supancic III, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4325-4333
We tackle the problem of estimating the 3D pose of an individual's upper limbs (arms+hands) from a chest mounted depth-camera. Importantly, we consider pose estimation during everyday interactions with objects. Past work shows that strong pose+viewpoint priors and depth-based features are crucial for robust performance. In egocentric views, hands and arms are observable within a well defined volume in front of the camera. We call this volume an egocentric workspace. A notable property is that hand appearance correlates with workspace location. To exploit this correlation, we classify arm+hand configurations in a global egocentric coordinate frame, rather than a local scanning window. This greatly simplify the architecture and improves performance. We propose an efficient pipeline which 1) generates synthetic workspace exemplars for training using a virtual chest-mounted camera whose intrinsic parameters match our physical camera, 2) computes perspective-aware depth features on this entire volume and 3) recognizes discrete arm+hand pose classes through a sparse multi-class SVM. We achieve state-of-the-art hand pose recognition performance from egocentric RGB-D images in real-time.
****
Simultaneous Time-of-Flight Sensing and Photometric Stereo With a Single ToF Sensor
Changpeng Ti, Ruigang Yang, James Davis, Zhigeng Pan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4334-4342
We present a novel system which incorporates photometric stereo with the Time-of-Flight depth sensor. Adding to the classic ToF, the system utilizes multiple point light sources that enable the capturing of a normal field whilst taking depth images. Two calibration methods are proposed to determine the light sources' positions given the ToF sensor's relatively low resolution. An iterative refinement algorithm is formulated to account for the extra phase delays caused by the positions of the light sources. We find in experiments that the system is comparable to the classic ToF in depth accuracy, and it is able to recover finer details that are lost due to the noise level of the ToF sensor.
****
Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances
Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, Joachim Denzler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4343-4352
Current visual recognition algorithms are "hungry" for data but massive annotation is extremely costly. Therefore, active learning algorithms are required that reduce labeling efforts to a minimum by selecting examples that are most valuable for labeling. In active learning, all categories occurring in collected data are usually assumed to be known in advance and experts should be able to label every requested instance. But do these assumptions really hold in practice? Could you name all categories in every image? Existing algorithms completely ignore the fact that there are certain examples where an oracle can not provide an answer or which even do not belong to the current problem domain. Ideally, active learning techniques should be able to discover new classes and at the same time cope with queries an expert is not able or willing to label. To meet these observations, we present a variant of the expected model output change principle for active learning and discovery in the presence of unnameable instances. Our experiments show that in these realistic scenarios, our approach substantially outperforms previous active learning methods, which are often not even able to improve with respect to the baseline of random query selection.
****
Learning to Compare Image Patches via Convolutional Neural Networks
Sergey Zagoruyko, Nikos Komodakis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4353-4361
In this paper we show how to learn directly from image data (i.e., without resorting to manually-designed features) a general similarity function for comparing

image patches, which is a task of fundamental importance for many computer visio n problems. To encode such a function, we opt for a CNN-based model that is tra ined to account for a wide variety of changes in image appearance. To that end, we explore and study multiple neural network architectures, which are specifical ly adapted to this task. We show that such an approach can significantly outperf orm the state-of-the-art on several problems and benchmark datasets.
****

Watch-n-Patch: Unsupervised Understanding of Actions and Relations
Chenxia Wu, Jiemi Zhang, Silvio Savarese, Ashutosh Saxena; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4362-4370
We focus on modeling human activities comprising multiple actions in a completel y unsupervised setting. Our model learns the high-level action co-occurrence and temporal relations between the actions in the activity video. We consider the v ideo as a sequence of short-term action clips, called action-words, and an activ ity is about a set of action-topics indicating which actions are present in the video. Then we propose a new probabilistic model relating the action-words and t he action-topics. It allows us to model long-range action relations that commonl y exist in the complex activity, which is challenging to capture in the previous works. We apply our model to unsupervised action segmentation and recognition, and also to a novel application that detects forgotten actions, which we call ac tion patching. For evaluation, we also contribute a new challenging RGB-D activ ity video dataset recorded by the new Kinect v2, which contains several human da ily activities as compositions of multiple actions interacted with different obj ects. The extensive experiments show the effectiveness of our model.
****

Optimal Graph Learning With Partial Tags and Multiple Features for Image and Vid eo Annotation
Lianli Gao, Jingkuan Song, Feiping Nie, Yan Yan, Nicu Sebe, Heng Tao Shen; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4371-4379
In multimedia annotation, due to the time constraints and the tediousness of man ual tagging, it is quite common to utilize both tagged and untagged data to impr ove the performance of supervised learning when only limited tagged training dat a are available. This is often done by adding a geometrically based regulari zation term in the objective function of a supervised learning model. In this ca se, a similarity graph is indispensable to exploit the geometrical relationships among the training data points, and the graph construction scheme essentially d etermines the performance of these graph-based learning algorithms. However, mos t of the existing works construct the graph empirically and are usually based on a single feature without using the label information. In this paper, we propose a semi-supervised annotation approach by learning an optimal graph (OGL) from m ulti-cues (i.e., partial tags and multiple features) which can more accurately e mbed the relationships among the data points. We further extend our model to add ress out-of-sample and noisy label issues. Extensive experiments on four publ ic datasets show the consistent superiority of OGL over state-of-the-art methods by up to 12% in terms of mean average precision.
****

DeepEdge: A Multi-Scale Bifurcated Deep Network for Top-Down Contour Detection
Gedas Bertasius, Jianbo Shi, Lorenzo Torresani; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4380-4389
Contour detection has been a fundamental component in many image segmentation an d object detection systems. Most previous work utilizes low-level features such as texture or saliency to detect contours and then use them as cues for a higher -level task such as object detection. However, we claim that recognizing objects and predicting contours are two mutually related tasks. Contrary to traditional approaches, we show that we can invert the commonly established pipeline: inste ad of detecting contours with low-level cues for a higher-level recognition task , we exploit object-related features as high-level cues for contour detection. We achieve this goal by means of a multi-scale deep network that consists of fi

ve convolutional layers and a bifurcated fully-connected sub-network. The sectio n from the input layer to the fifth convolutional layer is fixed and directly li fted from a pre-trained network optimized over a large-scale object classificati on task. This section of the network is applied to four different scales of the image input. These four parallel and identical streams are then attached to a bi furcated sub-network consisting of two independently-trained branches. One branc h learns to predict the contour likelihood (with a classification objective) whe reas the other branch is trained to learn the fraction of human labelers agreein g about the contour presence at a given point (with a regression criterion). W e show that without any feature engineering our multi-scale deep learning approa ch achieves state-of-the-art results in contour detection.
****

Picture: A Probabilistic Programming Language for Scene Perception
Tejas D. Kulkarni, Pushmeet Kohli, Joshua B. Tenenbaum, Vikash Mansinghka; Proce edings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4390-4399
Recent progress on probabilistic modeling and statistical learning, coupled with the availability of large training datasets, has led to remarkable progress in computer vision. Generative probabilistic models, or analysis-by-synthesis appro aches, can capture rich scene structure but have been less widely applied than t heir discriminative counterparts, as they often require considerable problem-spe cific engineering in modeling and inference, and inference is typically seen as requiring slow, hypothesize-and-test Monte Carlo methods. Here we present Pictu re, a probabilistic programming language for scene understanding that allows res earchers to express complex generative vision models, while automatically solvin g them using fast general-purpose inference machinery. Picture provides a stocha stic scene language that can express generative models for arbitrary 2D/3D scene s, as well as a hierarchy of representation layers for comparing scene hypothese s with observed images by matching not simply pixels, but also more abstract fea tures (e.g., contours, deep neural network activations). Inference can flexibly integrate advanced Monte Carlo strategies with fast bottom-up data-driven method s. Thus both representations and inference strategies can build directly on prog ress in discriminatively trained systems to make generative vision more robust a nd efficient. We use Picture to write programs for 3D face analysis, 3D human po se estimation, and 3D object reconstruction - each competitive with specially en gineered baselines.
****

Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization
Julien Valentin, Matthias Niessner, Jamie Shotton, Andrew Fitzgibbon, Shahram Iz adi, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision an d Pattern Recognition (CVPR), 2015, pp. 4400-4408
Recent advances in camera relocalization use predictions from a regression fores t to guide the camera pose optimization procedure. In these methods, each tree a ssociates one pixel with a point in the scene's 3D world coordinate frame. In pr evious work, these predictions were point estimates and the subsequent camera po se optimization implicitly assumed an isotropic distribution of these estimates. In this paper, we train a regression forest to predict mixtures of anisotropic 3D Gaussians and show how the predicted uncertainties can be taken into account for continuous pose optimization. Experiments show that our proposed method is a ble to relocalize up to 40% more frames than the state-of-the-art.
****

Fusing Subcategory Probabilities for Texture Classification
Yang Song, Weidong Cai, Qing Li, Fan Zhang, David Dagan Feng, Heng Huang; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4409-4417
Texture, as a fundamental characteristic of objects, has attracted much attentio n in computer vision research. Performance of texture classification is however still lacking for some challenging cases, largely due to the high intra-class va riation and low inter-class distinction. To tackle these issues, in this paper, we propose a sub-categorization model for texture classification. By clustering

each class into subcategories, classification probabilities at the subcategory-level are computed based on between-subcategory distinctiveness and within-subcategory representativeness. These subcategory probabilities are then fused based on their contribution levels and cluster qualities. This fused probability is added to the multiclass classification probability to obtain the final class label. Our method was applied to texture classification on three challenging datasets - KTH-TIPS2, FMD and DTD, and has shown excellent performance in comparison with the state-of-the-art approaches.
****

Video Event Recognition With Deep Hierarchical Context Model
Xiaoyang Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4418-4427
Video event recognition still faces great challenges due to large intra-class variation and low image resolution, in particular for surveillance videos. To mitigate these challenges and to improve the event recognition performance, various context information from the feature level, the semantic level, as well as the prior level is utilized. Different from most existing context approaches that utilize context in one of the three levels through shallow models like support vector machines, or probabilistic models like BN and MRF, we propose a deep hierarchical context model that simultaneously learns and integrates context at all three levels, and holistically utilizes the integrated contexts for event recognition. We first introduce two types of context features describing the event neighborhood, and then utilize the proposed deep model to learn the middle level representations and combine the bottom feature level, middle semantic level and top prior level contexts together for event recognition. The experiments on state of art surveillance video event benchmarks including VIRAT 1.0 Ground Dataset, VIRAT 2.0 Ground Dataset, and the UT-Interaction Dataset demonstrate that the proposed model is quite effective in utilizing the context information for event recognition. It outperforms the existing context approaches that also utilize multiple level contexts on these event benchmarks.
****

Object-Based RGBD Image Co-Segmentation With Mutex Constraint
Huazhu Fu, Dong Xu, Stephen Lin, Jiang Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4428-4436
We present an object-based co-segmentation method that takes advantage of depth data and is able to correctly handle noisy images in which the common foreground object is missing. With RGBD images, our method utilizes the depth channel to enhance identification of similar foreground objects via a proposed RGBD co-saliency map, as well as to improve detection of object-like regions and provide depth-based local features for region comparison. To accurately deal with noisy images where the common object appears more than or less than once, we formulate co-segmentation in a fully-connected graph structure together with mutual exclusion (mutex) constraints that prevent improper solutions. Experiments show that this object-based RGBD co-segmentation with mutex constraints outperforms related techniques on an RGBD co-segmentation dataset, while effectively processing noisy images. Moreover, we show that this method also provides performance comparable to state-of-the-art RGB co-segmentation techniques on regular RGB images with depth maps estimated from them.
****

Associating Neural Word Embeddings With Deep Image Representations Using Fisher Vectors
Benjamin Klein, Guy Lev, Gil Sadeh, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4437-4446
In recent years, the problem of associating a sentence with an image has gained a lot of attention. This work continues to push the envelope and makes further progress in the performance of image annotation and image search by a sentence tasks. In this work, we are using the Fisher Vector as a sentence representation by pooling the word2vec embedding of each word in the sentence. The Fisher Vector is typically taken as the gradients of the log-likelihood of descriptors, with respect to the parameters of a Gaussian Mixture Model (GMM). In this work we pre

sent two other Mixture Models and derive their Expectation-Maximization and Fisher Vector expressions. The first is a Laplacian Mixture Model (LMM), which is based on the Laplacian distribution. The second Mixture Model presented is a Hybrid Gaussian-Laplacian Mixture Model (HGLMM) which is based on a weighted geometric mean of the Gaussian and Laplacian distribution. Finally, by using the new Fisher Vectors derived from HGLMMs to represent sentences, we achieve state-of-the-art results for both the image annotation and the image search by a sentence tasks on four benchmarks: Pascal1K, Flickr8K, Flickr30K, and COCO.
****

3D Shape Estimation From 2D Landmarks: A Convex Relaxation Approach
Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4447-4455
We investigate the problem of estimating the 3D shape of an object, given a set of 2D landmarks in a single image. To alleviate the reconstruction ambiguity, a widely-used approach is to confine the unknown 3D shape within a shape space built upon existing shapes. While this approach has proven to be successful in various applications, a challenging issue remains, i.e., the joint estimation of shape parameters and camera-pose parameters requires to solve a nonconvex optimization problem. The existing methods often adopt an alternating minimization scheme to locally update the parameters, and consequently the solution is sensitive to initialization. In this paper, we propose a convex formulation to address this problem and develop an efficient algorithm to solve the proposed convex program. We demonstrate the exact recovery property of the proposed method, its merits compared to alternative methods, and the applicability in human pose and car shape estimation.
****

3D All The Way: Semantic Segmentation of Urban Scenes From Start to End in 3D
Andelo Martinovic, Jan Knopp, Hayko Riemenschneider, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4456-4465
We propose a new approach for semantic segmentation of 3D city models. Starting from an SfM reconstruction of a street-side scene, we perform classification and facade splitting purely in 3D, obviating the need for slow image-based semantic segmentation methods. We show that a properly trained pure-3D approach produces high quality labelings, with significant speed benefits (20x faster) allowing us to analyze entire streets in a matter of minutes. Additionally, if speed is not of the essence, the 3D labeling can be combined with the results of a state-of-the-art 2D classifier, further boosting the performance. Further, we propose a novel facade separation based on semantic nuances between facades. Finally, inspired by the use of architectural principles for 2D facade labeling, we propose new 3D-specific principles and an efficient optimization scheme based on an integer quadratic programming formulation.
****

Fast Bilateral-Space Stereo for Synthetic Defocus
Jonathan T. Barron, Andrew Adams, YiChang Shih, Carlos Hernandez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4466-4474
Given a stereo pair it is possible to recover a depth map and use that depth to render a synthetically defocused image. Though stereo algorithms are well-studied, rarely are those algorithms considered solely in the context of producing these defocused renderings. In this paper we present a technique for efficiently producing disparity maps using a novel optimization framework in which inference is performed in "bilateral-space". Our approach produces higher-quality "defocus" results than other stereo algorithms while also being 10-100 times faster than comparable techniques.
****

Large-Scale and Drift-Free Surface Reconstruction Using Online Subvolume Registration
Nicola Fioraio, Jonathan Taylor, Andrew Fitzgibbon, Luigi Di Stefano, Shahram Iz

adi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2015, pp. 4475-4483
Depth cameras have helped commoditize 3D digitization of the real-world. It is n
ow feasible to use a single Kinect-like camera to scan in an entire building or
other large-scale scenes. At large scales, however, there is an inherent challen
ge of dealing with distortions and drift due to accumulated pose estimation erro
rs. Existing techniques suffer from one or more of the following: a) requiring a
n expensive offline global optimization step taking hours to compute; b) needing
 a full second pass over the input depth frames to correct for accumulated error
s; c) relying on RGB data alongside depth data to optimize poses; or d) requirin
g the user to create explicit loop closures to allow gross alignment errors to b
e resolved. In this paper, we present a method that addresses all of these issue
s. Our method supports online model correction, without needing to reprocess or
store any input depth data. Even while performing global correction of a large 3
D model, our method takes only minutes rather than hours to compute. Our model d
oes not require any explicit loop closures to be detected and, finally, relies o
n depth data alone, allowing operation in low-lighting conditions. We show quali
tative results on many large scale scenes, highlighting the lack of error and dr
ift in our reconstructions. We compare to state of the art techniques and demons
trate large-scale dense surface reconstruction "in the dark", a capability not o
ffered by RGB-D techniques.
****

Fast Randomized Singular Value Thresholding for Nuclear Norm Minimization
Tae-Hyun Oh, Yasuyuki Matsushita, Yu-Wing Tai, In So Kweon; Proceedings of the I
EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4484
-4493
Rank minimization problem can be boiled down to either Nuclear Norm Minimization
 (NNM) or Weighted NNM (WNNM) problem. The problems related to NNM (or WNNM) can
 be solved iteratively by applying a closed-form proximal operator, called Singu
lar Value Thresholding (SVT) (or Weighted SVT), but they suffer from high comput
ational cost to compute a Singular Value Decomposition (SVD) at each iteration.
In this paper, we propose an accurate and fast approximation method for SVT, cal
led fast randomized SVT (FRSVT), where we avoid direct computation of SVD. The k
ey idea is to extract an approximate basis for the range of a matrix from its co
mpressed matrix. Given the basis, we compute the partial singular values of the
original matrix from a small factored matrix. While the basis approximation is t
he bottleneck, our method is already severalfold faster than thin SVD. By adopti
ng a range propagation technique, we can further avoid one of the bottleneck at
each iteration. Our theoretical analysis provides a stepping stone between the a
pproximation bound of SVD and its effect to NNM via SVT. Along with the analysis
, our empirical results on both quantitative and qualitative studies show our ap
proximation rarely harms the convergence behavior of the host algorithms. We app
ly it and validate the efficiency of our method on various vision problems, e.g.
 subspace clustering, weather artifact removal, simultaneous multi-image alignme
nt and rectification.
****

LMI-Based 2D-3D Registration: From Uncalibrated Images to Euclidean Scene
Danda Pani Paudel, Adlane Habed, Cedric Demonceaux, Pascal Vasseur; Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015,
pp. 4494-4502
This paper investigates the problem of registering a scanned scene, represented
by 3D Euclidean point coordinates, and two or more uncalibrated cameras. An unkn
own subset of the scanned points have their image projections detected and match
ed across images. The proposed approach assumes the cameras only known in some a
rbitrary projective frame and no calibration or autocalibration is required. The
 devised solution is based on a Linear Matrix Inequality (LMI) framework that al
lows simultaneously estimating the projective transformation relating the camera
s to the scene and establishing 2D-3D correspondences without triangulating imag
e points. The proposed LMI framework allows both deriving triangulation-free LMI
 cheirality conditions and establishing putative correspondences between 3D volu

mes (boxes) and 2D pixel coordinates. Two registration algorithms, one exploitin g the scene's structure and the other concerned with robustness, are presented. Both algorithms employ the Branch-and-Prune paradigm and guarantee convergence t o a global solution under mild initial bound conditions. The results of our expe riments are presented and compared against other approaches.
****

Clique-Graph Matching by Preserving Global & Local Structure
Wei-Zhi Nie, An-An Liu, Zan Gao, Yu-Ting Su; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4503-4510
This paper originally proposes the clique-graph and further presents a clique-gr aph matching method by preserving global and local structures. Especially, we fo rmulate the objective function of clique-graph matching with respective to two l atent variables, the clique information in the original graph and the pairwise c lique correspondence constrained by the one-to-one matching. Since the objective function is not jointly convex to both latent variables, we decompose it into t wo consecutive steps for optimization: 1) clique-to-clique similarity measure by preserving local unary and pairwise correspondences; 2) graph-to-graph similari ty measure by preserving global clique-to-clique correspondence. Extensive exper iments on the synthetic data and real images show that the proposed method can o utperform representative methods especially when both noise and outliers exist.
****

Appearance-Based Gaze Estimation in the Wild
Xucong Zhang, Yusuke Sugano, Mario Fritz, Andreas Bulling; Proceedings of the IE EE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4511-4520
Appearance-based gaze estimation is believed to work well in real-world settings , but existing datasets have been collected under controlled laboratory conditio ns and methods have been not evaluated across multiple datasets. In this work we study appearance-based gaze estimation in the wild. We present the MPIIGaze dat aset that contains 213,659 images we collected from 15 participants during natur al everyday laptop use over more than three months. Our dataset is significantly more variable than existing ones with respect to appearance and illumination. W e also present a method for in-the-wild appearance-based gaze estimation using m ultimodal convolutional neural networks that significantly outperforms state-of- the art methods in the most challenging cross-dataset evaluation. We present an extensive evaluation of several state-of-the-art image-based gaze estimation alg orithms on three current datasets, including our own. This evaluation provides c lear insights and allows us to identify key research challenges of gaze estimati on in the wild.
****

One-Day Outdoor Photometric Stereo via Skylight Estimation
Jiyoung Jung, Joon-Young Lee, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4521-4529
We present an outdoor photometric stereo method using images captured in a singl e day. We simulate a sky hemisphere for each image according to its GPS and time stamp, and parameterize the obtained sky hemisphere into a quadratic skylight an d a Gaussian sunlight distribution. Unlike previous works which usually model ou tdoor illumination as a sum of constant ambient light and a distant point light, our method models natural illumination according to a popular sky model and thu s provides sufficient constraints for shape reconstruction from one day images. We generate pixel profiles of uniformly sampled unit vectors for the correspondi ng time of captures and evaluate them using correlation with the actual pixel pr ofiles. The estimated surface normal is refined by MRF optimization. We have tes ted our method to recover objects and scenes of various sizes in real-world outd oor daylight.
****

A New Retraction for Accelerating the Riemannian Three-Factor Low-Rank Matrix Co mpletion Algorithm
Zhizhong Li, Deli Zhao, Zhouchen Lin, Edward Y. Chang; Proceedings of the IEEE C onference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4530-4538

The Riemannian three-factor matrix completion (R3MC) algorithm is one of the sta
te-of-the-art geometric optimization methods for the low-rank matrix completion
problem. It is a nonlinear conjugate-gradient method optimizing on a quotient Ri
emannian manifold. In the line search step, R3MC approximates the minimum point
on the searching curve by minimizing on the line tangent to the curve. However,
finding the exact minimum point by iteration is too expensive. We address this i
ssue by proposing a new retrac with a minimizing property. This special property
 provides the exact minimization for the line search by establishing corresponde
nces between points on the searching curve and points on the tangent line. Accel
erated R3MC, which is R3MC equipped with this new retraction, outperforms the or
iginal algorithm and other geometric algorithms for matrix completion in our emp
irical study.
****

Heteroscedastic Max-Min Distance Analysis
Bing Su, Xiaoqing Ding, Changsong Liu, Ying Wu; Proceedings of the IEEE Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4539-4547
Many discriminant analysis methods such as LDA and HLDA actually maximize the av
erage pairwise distances between classes, which often causes the class separatio
n problem. Max-min distance analysis (MMDA) addresses this problem by maximizing
 the minimum pairwise distance in the latent subspace, but it is developed under
 the homoscedastic assumption. This paper proposes Heteroscedastic MMDA (HMMDA)
methods that explore the discriminative information in the difference of intra-c
lass scatters for dimensionality reduction. WHMMDA maximizes the minimal pairwis
e Chenoff distance in the whitened space. OHMMDA incorporates this objective and
 the minimization of class compactness into a trace quotient formulation and imp
oses an orthogonal constraint to the final transformation, which can be solved b
y a bisection search algorithm. Two variants of OHMMDA are further proposed to e
ncode the margin information. Experiments on several UCI Machine Learning datase
ts and the Yale Face database demonstrate the effectiveness of the proposed HMMD
A methods.
****

Sparse Composite Quantization
Ting Zhang, Guo-Jun Qi, Jinhui Tang, Jingdong Wang; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4548-4556
The quantization techniques have shown competitive performance in approximate ne
arest neighbor search. The state-of-the-art algorithm, composite quantization, t
akes advantage of the compositionabity, i.e., the vector approximation accuracy,
 as opposed to product quantization and Cartesian k-means. However, we have obse
rved that the runtime cost of computing the distance table in composite quantiza
tion, which is used as a lookup table for fast distance computation, becomes non
negligible in real applications, e.g., reordering the candidates retrieved from
the inverted index when handling very large scale databases. To address this pro
blem, we develop a novel approach, called sparse composite quantization, which c
onstructs sparse dictionaries. The benefit is that the distance evaluation betwe
en the query and the dictionary element (a sparse vector) is accelerated using t
he efficient sparse vector operation, and thus the cost of distance table comput
ation is reduced a lot. Experiment results on large scale ANN retrieval tasks (1
M SIFTs and 1B SIFTs) and applications to object retrieval show that the propose
d approach yields competitive performance: superior search accuracy to product q
uantization and Cartesian k-means with almost the same computing cost, and much
faster ANN search than composite quantization with the same level of accuracy.
****

Sparse Representation Classification With Manifold Constraints Transfer
Baochang Zhang, Alessandro Perina, Vittorio Murino, Alessio Del Bue; Proceedings
 of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015,
 pp. 4557-4565
The fact that image data samples lie on a manifold has been successfully exploit
ed in many learning and inference problems. In this paper we leverage the specif
ic structure of data in order to improve recognition accuracies in general recog
nition tasks. In particular we propose a novel framework that allows to embed ma

nifold priors into sparse representation-based classification (SRC) approaches. We also show that manifold constraints can be transferred from the data to the optimized variables if these are linearly correlated. Using this new insight, we define an efficient alternating direction method of multipliers (ADMM) that can consistently integrate the manifold constraints during the optimization process. This is based on the property that we can recast the problem as the projection over the manifold via a linear embedding method based on the Geodesic distance. The proposed approach is successfully applied on face, digit, action and objects recognition showing a consistently increase on performance when compared to the state of the art.
****

CIDEr: Consensus-Based Image Description Evaluation
Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566-4575
Automatically describing an image with a sentence is a long-standing challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition, etc., there is renewed interest in this area. However, evaluating the quality of descriptions has proven to be challenging. We propose a novel paradigm for evaluating image descriptions that uses human consensus. This paradigm consists of three main parts: a new triplet-based method of collecting human annotations to measure consensus, a new automated metric that captures consensus, and two new datasets: PASCAL-50S and ABSTRACT-50S that contain 50 sentences describing each image. Our simple metric captures human judgment of consensus better than existing metrics across sentences generated by various sources. We also evaluate five state-of-the-art image description approaches using this new protocol and provide a benchmark for future comparisons. A version of CIDEr named CIDEr-D is available as a part of MS COCO evaluation server to enable systematic evaluation and benchmarking.
****

Joint Inference of Groups, Events and Human Roles in Aerial Videos
Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, Song Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4576-4584
With the advent of drones, aerial video analysis becomes increasingly important; yet, it has received scant attention in the literature. This paper addresses a new problem of parsing low-resolution aerial videos of large spatial areas, in terms of 1) grouping, 2) recognizing events and 3) assigning roles to people engaged in events. We propose a novel framework aimed at conducting joint inference of the above tasks, as reasoning about each in isolation typically fails in our setting. Given noisy tracklets of people and detections of large objects and scene surfaces (e.g., building, grass), we use a spatiotemporal AND-OR graph to drive our joint inference, using Markov Chain Monte Carlo and dynamic programming. We also introduce a new formalism of spatiotemporal templates characterizing latent sub-events. For evaluation, we have collected and released a new aerial videos dataset using a hex-rotor flying over picnic areas rich with group events. Our results demonstrate that we successfully address above inference tasks under challenging conditions.
****

Photometric Stereo With Near Point Lighting: A Solution by Mesh Deformation
Wuyuan Xie, Chengkai Dai, Charlie C. L. Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4585-4593
We tackle the problem of photometric stereo under near point lighting in this paper. Different from the conventional formulation of photometric stereo that assumes parallel lighting, photometric stereo under the near point lighting condition is a nonlinear problem as the local surface normals are coupled with its distance to the camera as well as the light sources. To solve this non-linear problem of PS with near point lighting, a local/global mesh deformation approach is developed in our work to determine the position and the orientation of a facet simultaneously, where each facet is corresponding to a pixel in the image captured b

y the camera. Unlike nonlinear optimization schemes, the mesh deformation in our approach is decoupled into an iteration of interlaced steps of local projection and global blending. Experimental results verify that our method can generate accurate estimation of surface shape under near point lighting in a few iterations. Besides, this approach is robust to errors on the positions of light sources and is easy to be implemented.
****

Efficient Label Collection for Unlabeled Image Datasets
Maggie Wigness, Bruce A. Draper, J. Ross Beveridge; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4594-4602
Visual classifiers are part of many applications including surveillance, autonomous navigation and scene understanding. The raw data used to train these classifiers is abundant and easy to collect but lacks labels. Labels are necessary for training supervised classifiers, but the labeling process requires significant human effort. Techniques like active learning and group-based labeling have emerged to help reduce the labeling workload. However, the possibility of collecting label noise affects either the efficiency of these systems or the performance of the trained classifiers. Further, many introduce latency by iteratively re-training classifiers or re-clustering data. We introduce a technique that searches for structural change in hierarchically clustered data to identify a set of clusters that span a spectrum of visual concept granularities. This allows us to efficiently label clusters with less label noise and produce high performing classifiers. The data is hierarchically clustered only once, eliminating latency during the labeling process. Using benchmark data we show that collecting labels with our approach is more efficient than existing labeling techniques, and achieves higher classification accuracy. Finally, we demonstrate the speed and efficiency of our system using real-world data collected for an autonomous navigation task.
****

Separating Objects and Clutter in Indoor Scenes
Salman H. Khan, Xuming He, Mohammed Bennamoun, Ferdous Sohel, Roberto Togneri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4603-4611
Objects' spatial layout estimation and clutter identification are two important tasks to understand indoor scenes. We propose to solve both of these problems in a joint framework using RGBD images of indoor scenes. In contrast to recent approaches which focus on either one of these two problems, we perform `fine grained structure categorization' by predicting all the major objects and simultaneously labeling the cluttered regions. A conditional random field model is proposed to incorporate a rich set of local appearance, geometric features and interactions between the scene elements. We take a structural learning approach with a loss of 3D localisation to estimate the model parameters from a large annotated RGBD dataset, and a mixed integer linear programming formulation for inference. We demonstrate that our approach is able to detect cuboids and estimate cluttered regions across many different object and scene categories in the presence of occlusion, illumination and appearance variations.
****

FaLRR: A Fast Low Rank Representation Solver
Shijie Xiao, Wen Li, Dong Xu, Dacheng Tao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4612-4620
Low rank representation (LRR) has shown promising performance for various computer vision applications such as face clustering. Existing algorithms for solving LRR usually depend on its two-variable formulation which contains the original data matrix. In this paper, we develop a fast LRR solver called FaLRR, by reformulating LRR as a new optimization problem with regard to factorized data (which is obtained by skinny SVD of the original data matrix). The new formulation benefits the corresponding optimization and theoretical analysis. Specifically, to solve the resultant optimization problem, we propose a new algorithm which is not only efficient but also theoretically guaranteed to obtain a globally optimal solution. Regarding the theoretical analysis, the new formulation is helpful for deriving some interesting properties of LRR. Last but not least, the proposed alg

orithm can be readily incorporated into an existing distributed framework of LRR for further acceleration. Extensive experiments on synthetic and real-world datasets demonstrate that our FaLRR achieves order-of-magnitude speedup over existing LRR solvers, and the efficiency can be further improved by incorporating our algorithm into the distributed framework of LRR.

****

Simulating Makeup Through Physics-Based Manipulation of Intrinsic Image Layers

Chen Li, Kun Zhou, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4621-4629

We present a method for simulating makeup in a face image. To generate realistic results without detailed geometric and reflectance measurements of the user, we propose to separate the image into intrinsic image layers and alter them according to proposed adaptations of physically-based reflectance models. Through this layer manipulation, the measured properties of cosmetic products are applied while preserving the appearance characteristics and lighting conditions of the target face. This approach is demonstrated on various forms of cosmetics including foundation, blush, lipstick, and eye shadow. Experimental results exhibit a close approximation to ground truth images, without artifacts such as transferred personal features and lighting effects that degrade the results of image-based makeup transfer methods.

****

Correlation Filters With Limited Boundaries

Hamed Kiani Galoogahi, Terence Sim, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4630-4638

Correlation filters take advantage of specific properties in the Fourier domain allowing them to be estimated efficiently: O(ND log D) in the frequency domain, versus O(D^3 + ND^2) spatially where D is signal length, and N is the number of signals. Recent extensions to correlation filters, such as MOSSE, have reignited interest of their use in the vision community due to their robustness and attractive computational properties. In this paper we demonstrate, however, that this computational efficiency comes at a cost. Specifically, we demonstrate that only 1/D proportion of shifted examples are unaffected by boundary effects which has a dramatic effect on detection/tracking performance. In this paper, we propose a novel approach to correlation filter estimation that: (i) takes advantage of inherent computational redundancies in the frequency domain, (ii) dramatically reduces boundary effects, and (iii) is able to implicitly exploit all possible patches densely extracted from training examples during learning process. Impressive object tracking and detection results are presented in terms of both accuracy and computational efficiency.

****

Shape-Based Automatic Detection of a Large Number of 3D Facial Landmarks

Syed Zulqarnain Gilani, Faisal Shafait, Ajmal Mian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4639-4648

We present an algorithm for automatic detection of a large number of anthropometric landmarks on 3D faces. Our approach does not use texture and is completely shape based in order to detect landmarks that are morphologically significant. The proposed algorithm evolves level set curves with adaptive geometric speed functions to automatically extract effective seed points for dense correspondence. Correspondences are established by minimizing the bending energy between patches around seed points of given faces to those of a reference face. Given its hierarchical structure, our algorithm is capable of establishing thousands of correspondences between a large number of faces. Finally, a morphable model based on the dense corresponding points is fitted to an unseen query face for transfer of correspondences and hence automatic detection of landmarks. The proposed algorithm can detect any number of pre-defined landmarks including subtle landmarks that are even difficult to detect manually. Extensive experimental comparison on two benchmark databases containing 6,507 scans shows that our algorithm outperforms six state of the art algorithms.

****

Material Classification With Thermal Imagery

Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Chandra Kambhamettu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4649-4656

Material classification is an important area of research in computer vision. Typical algorithms use color and texture information for classification, but there are problems due to varying lighting conditions and diversity of colors in a single material class. In this work we study the use of long wave infrared (i.e. thermal) imagery for material classification. Thermal imagery has the benefit of relative invariance to color changes, invariance to lighting conditions, and can even work in the dark. We collect a database of 21 different material classes with both color and thermal imagery. We develop a set of features that describe water permeation and heating/cooling properties, and test several variations on these methods to obtain our final classifier. The results show that the proposed method outperforms typical color and texture features, and when combined with color information, the results are improved further.

****

Deeply Learned Attributes for Crowded Scene Understanding

Jing Shao, Kai Kang, Chen Change Loy, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4657-4666

Crowded scene understanding is a fundamental problem in computer vision. In this study, we develop a multi-task deep model to jointly learn and combine appearance and motion features for crowd understanding. We propose crowd motion channels as the input of the deep model and the channel design is inspired by generic properties of crowd systems. To well demonstrate our deep model, we construct a new large-scale WWW Crowd dataset with 10000 videos from 8257 crowded scenes, and build an attribute set with 94 attributes on WWW. We further measure user study performance on WWW and compare this with the proposed deep models. Extensive experiments show that our deep models display significant performance improvements in cross-scene attribute recognition compared to strong crowd-related feature-based baselines, and the deeply learned features behave an superior performance in multi-task learning.

****

Learning To Look Up: Realtime Monocular Gaze Correction Using Machine Learning

Daniil Kononenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4667-4675

We revisit the well-known problem of gaze correction and present a solution based on supervised machine learning. At training time, our system observes pairs of images, where each pair contains the face of the same person with a fixed angular difference in gaze direction. It then learns to synthesize the second image of a pair from the first one. After learning, the system becomes able to redirect the gaze of a previously unseen person by the same angular difference. Unlike many previous solutions to gaze problem in videoconferencing, ours is purely monocular, i.e. it does not require any hardware apart from an in-built web-camera of a laptop. We base our machine learning implementation on a special kind of decision forests that predict a displacement (flow) vector for each pixel in the input image. As a result, our system is highly efficient (runs in real-time on a single core of a modern laptop). In the paper, we demonstrate results on a variety of videoconferencing frames and evaluate the method quantitatively on the held-out set of registered images. The supplementary video shows example sessions of our system at work.

****

Background Subtraction via Generalized Fused Lasso Foreground Modeling

Bo Xin, Yuan Tian, Yizhou Wang, Wen Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4676-4684

Background Subtraction (BS) is one of the key steps in video analysis. Many background models have been proposed and achieved promising performance on public data sets. However, due to challenges such as illumination change, dynamic background etc. the resulted foreground segmentation often consists of holes as well as background noise. In this regard, we consider generalized fused lasso regularization to quest for intact structured foregrounds. Together with certain assumpti

ons about the background, such as the low-rank assumption or the sparse composit ion assumption (depending on whether pure background frames are provided), we fo rmulate BS as a matrix decomposition problem using regularization terms for both the foreground and background matrices. Moreover, under the proposed formulatio n, the two generally distinctive background assumptions can be solved in a unifi ed manner. The optimization was carried out via applying the augmented Lagrange multiplier (ALM) method in such a way that a fast parametric-flow algorithm is u sed for updating the foreground matrix. Experimental results on several popular BS data sets demonstrate the advantage of the proposed model compared to state-o f-the-arts.
****

Mirror, Mirror on the Wall, Tell Me, Is the Error Small?
Heng Yang, Ioannis Patras; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4685-4693
Do object part localization methods produce bilaterally symmetric results on mir ror images? Surprisingly not, even though state of the art methods augment the t raining set with mirrored images. In this paper we take a closer look into this issue. We first introduce the concept of mirrorability as the ability of a model to produce symmetric results in mirrored images and introduce a corresponding m easure, namely the \textit{mirror error} that is defined as the difference betwe en the detection result on an image and the mirror of the detection result on i ts mirror image. We evaluate the mirrorability of several state of the art algor ithms in two of the most intensively studied problems, namely human pose estimat ion and face alignment. Our experiments lead to several interesting findings: 1) Most of state of the art methods struggle to preserve the mirror symmetry, desp ite the fact that they do have very similar overall performance on the original and mirror images; 2) the low mirrorability is not caused by training or testing sample bias - all algorithms are trained on both the original images and their mirrored versions; 3) the mirror error is strongly correlated to the localizatio n/alignment error (with correlation coefficients around 0.7). Since the mirror e rror is calculated without knowledge of the ground truth, we show two interestin g applications - in the first it is used to guide the selection of difficult sam ples and in the second to give feedback in a popular Cascaded Pose Regression me thod for face alignment.
****

Beyond Short Snippets: Deep Networks for Video Classification
Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici; Proceedings of the IEEE Conference on Computer Vis ion and Pattern Recognition (CVPR), 2015, pp. 4694-4702
Convolutional neural networks (CNNs) have been exten- sively applied for image r ecognition problems giving state- of-the-art results on recognition, detection, segmentation and retrieval. In this work we propose and evaluate several deep ne ural network architectures to combine image infor- mation across a video over lo nger time periods than previ- ously attempted. We propose two methods capable of han- dling full length videos. The first method explores various convolutional temporal feature pooling architectures, ex- amining the various design choices w hich need to be made when adapting a CNN for this task. The second proposed meth od explicitly models the video as an ordered sequence of frames. For this purpos e we employ a recurrent neural network that uses Long Short-Term Memory (LSTM) c ells which are connected to the output of the underlying CNN. Our best networks exhibit significant performance improve- ments over previously published results on the Sports 1 mil- lion dataset (73.1% vs. 60.9%) and the UCF-101 datasets wi th (88.2% vs. 87.9%) and without additional optical flow information (82.6% vs. 72.8%).
****

segDeepM: Exploiting Segmentation and Context in Deep Neural Networks for Object Detection
Yukun Zhu, Raquel Urtasun, Ruslan Salakhutdinov, Sanja Fidler; Proceedings of th e IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4 703-4711

In this paper, we propose an approach that exploits object segmentation in order to improve the accuracy of object detection. We frame the problem as inference in a Markov Random Field, in which each detection hypothesis scores object appearance as well as contextual information using Convolutional Neural Networks, and allows the hypothesis to choose and score a segment out of a large pool of accurate object segmentation proposals. This enables the detector to incorporate additional evidence when it is available and thus results in more accurate detections. Our experiments show an improvement of 4.1% in mAP over the R-CNN baseline on PASCAL VOC 2010, and 1.4% over the current state-of-the-art, demonstrating the power of our approach.
****

Situational Object Boundary Detection
Jasper R. R. Uijlings, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4712-4721
Intuitively, the appearance of true object boundaries varies from image to image. Hence the usual monolithic approach of training a single boundary predictor and applying it to all images regardless of their content is bound to be suboptimal. In this paper we therefore propose Situational Object Boundary Detection: we first define a variety of situations and train a specialized object boundary detector for each of them using [Dollar13]. Then given a test image, we classify it into these situations using its context, which we model by global image appearance. We apply the corresponding situational object boundary detectors, and fuse them based on the classification probabilities. In experiments on ImageNet [Russakovsky14], Microsoft COCO [Lin], and Pascal VOC 2012 segmentation [Everingham14] we show that our situational object boundary detection gives significant improvements over a monolithic approach. Additionally, our method substantially outperforms [Hariharan11iccv] on semantic contour detection on their SBD dataset.
****

Real-Time 3D Head Pose and Facial Landmark Estimation From Depth Images Using Triangular Surface Patch Features
Chavdar Papazov, Tim K. Marks, Michael Jones; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4722-4730
We present a real-time system for 3D head pose estimation and facial landmark localization using a commodity depth sensor. We introduce a novel triangular surface patch (TSP) descriptor, which encodes the shape of the 3D surface of the face within a triangular area. The proposed descriptor is viewpoint invariant, and it is robust to noise and to variations in the data resolution. Using a fast nearest neighbor lookup, TSP descriptors from an input depth map are matched to the most similar ones that were computed from synthetic head models in a training phase. The matched triangular surface patches in the training set are used to compute estimates of the 3D head pose and facial landmark positions in the input depth map. By sampling many TSP descriptors, many votes for pose and landmark positions are generated which together yield robust final estimates. We evaluate our approach on the publicly available Biwi Kinect Head Pose Database to compare it against state-of-the-art methods. Our results show a significant improvement in the accuracy of both pose and landmark location estimates while maintaining real-time speed.
****

Aligning 3D Models to RGB-D Images of Cluttered Scenes
Saurabh Gupta, Pablo Arbelaez, Ross Girshick, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4731-4740
The goal of this work is to represent objects in an RGB-D scene with corresponding 3D models from a library. We approach this problem by first detecting and segmenting object instances in the scene and then using a convolutional neural network (CNN) to predict the pose of the object. This CNN is trained using pixel surface normals in images containing renderings of synthetic objects. When tested on real data, our method outperforms alternative algorithms trained on real data. We then use this coarse pose estimate along with the inferred pixel support to align a small number of prototypical models to the data, and place into the scen

e the model that fits best. We observe a 48% relative improvement in performance at the task of 3D detection over the current state-of-the-art, while being an order of magnitude faster.

****

A Stable Multi-Scale Kernel for Topological Machine Learning

Jan Reininghaus, Stefan Huber, Ulrich Bauer, Roland Kwitt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4741-4748

Topological data analysis offers a rich source of valuable information to study vision problems. Yet, so far we lack a theoretically sound connection to popular kernel-based learning techniques, such as kernel SVMs or kernel PCA. In this work, we establish such a connection by designing a multi-scale kernel for persistence diagrams, a stable summary representation of topological features in data. We show that this kernel is positive definite and prove its stability with respect to the 1-Wasserstein distance. Experiments on two benchmark datasets for 3D shape classification/retrieval and texture recognition show considerable performance gains of the proposed method compared to an alternative approach that is based on the recently introduced persistence landscapes.

****

The Treasure Beneath Convolutional Layers: Cross-Convolutional-Layer Pooling for Image Classification

Lingqiao Liu, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4749-4757

A number of recent studies have shown that a Deep Convolutional Neural Network (DCNN) pretrained on a large dataset can be adopted as a universal image descriptor, and that doing so leads to impressive performance at a range of image classification tasks. Most of these studies, if not all, adopt activations of the fully-connected layer of a DCNN as the image or region representation and it is believed that convolutional layer activations are less discriminative. This paper, however, advocates that if used appropriately, convolutional layer activations constitute a powerful image representation. This is achieved by adopting a new technique proposed in this paper called cross-convolutional-layer pooling. More specifically, it extracts subarrays of feature maps of one convolutional layer as local features, and pools the extracted features with the guidance of the feature maps of the successive convolutional layer. Compared with existing methods that apply DCNNs in the similar local feature setting, the proposed method avoids the input image style mismatching issue which is usually encountered when applying fully connected layer activations to describe local regions. Also, the proposed method is easier to implement since it is codebook free and does not have any tuning parameters. By applying our method to four popular visual classification tasks, it is demonstrated that the proposed method can achieve comparable or in some cases significantly better performance than existing fully-connected layer based image representations.

****

Face Video Retrieval With Image Query via Hashing Across Euclidean Space and Riemannian Manifold

Yan Li, Ruiping Wang, Zhiwu Huang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4758-4767

Retrieving videos of a specific person given his/her face image as query becomes more and more appealing for applications like smart movie fast-forwards and suspect searching. It also forms an interesting but challenging computer vision task, as the visual data to match, i.e., still image and video clip are usually represented quite differently. Typically, face image is represented as point (i.e., vector) in Euclidean space, while video clip is seemingly modeled as a point (e.g., covariance matrix) on some particular Riemannian manifold in the light of its recent promising success. It thus incurs a new hashing-based retrieval problem of matching two heterogeneous representations, respectively in Euclidean space and Riemannian manifold. This work makes the first attempt to embed the two heterogeneous spaces into a common discriminant Hamming space. Specifically, we pro

pose Hashing across Euclidean space and Riemannian manifold (HER) by deriving a unified framework to firstly embed the two spaces into corresponding reproducing kernel Hilbert spaces, and then iteratively optimize the intra- and inter-space Hamming distances in a maxmargin framework to learn the hash functions for the two spaces. Extensive experiments demonstrate the impressive superiority of our method over the state-of-the-art competitive hash learning methods.
****

EgoSampling: Fast-Forward and Stereo for Egocentric Videos
Yair Poleg, Tavi Halperin, Chetan Arora, Shmuel Peleg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4768-4776
While egocentric cameras like GoPro are gaining popularity, the videos they capture are long, boring, and difficult to watch from start to end. Fast forwarding (i.e. frame sampling) is a natural choice for faster video browsing. However, this accentuates the shake caused by natural head motion, making the fast forwarded video useless. We propose EgoSampling, an adaptive frame sampling that gives more stable fast forwarded videos. Adaptive frame sampling is formulated as energy minimization, whose optimal solution can be found in polynomial time. In addition, egocentric video taken while walking suffers from the left-right movement of the head as the body weight shifts from one leg to another. We turn this drawback into a feature: Stereo video can be created by sampling the frames from the left most and right most head positions of each step, forming approximate stereo-pairs.
****

Social Saliency Prediction
Hyun Soo Park, Jianbo Shi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4777-4785
This paper presents a method to predict social saliency, the likelihood of joint attention, given an input image or video by leveraging the social interaction data captured by first person cameras. Inspired by electric dipole moments, we introduce a social formation feature that encodes the geometric relationship between joint attention and its social formation. We learn this feature from the first person social interaction data where we can precisely measure the locations of joint attention and its associated members in 3D. An ensemble classifier is trained to learn the geometric relationship. Using the trained classifier, we predict social saliency in real-world scenes with multiple social groups including scenes from team sports captured in a third person view. Our representation does not require directional measurements such as gaze directions. A geometric analysis of social interactions in terms of the F-formation theory is also presented.
****

Beyond Principal Components: Deep Boltzmann Machines for Face Modeling
Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Tien D. Bui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4786-4794
The "interpretation through synthesis", i.e. Active Appearance Models (AAMs) method, has received considerable attention over the past decades. It aims at "explaining" face images by synthesizing them via a parameterized model of appearance. It is quite challenging due to appearance variations of human face images, e.g. facial poses, occlusions, lighting, low resolution, etc. Since these variations are mostly non-linear, it is impossible to represent them in a linear model, such as Principal Component Analysis (PCA). This paper presents a novel Deep Appearance Models (DAMs) approach, an efficient replacement for AAMs, to accurately capture both shape and texture of face images under large variations. In this approach, three crucial components represented in hierarchical layers are modeled using the Deep Boltzmann Machines (DBM) to robustly capture the variations of facial shapes and appearances. DAMs are therefore superior to AAMs in inferring a representation for new face images under various challenging conditions. In addition, DAMs have ability to generate a compact set of parameters in higher level representation that can be used for classification, e.g. face recognition and facial age estimation. The proposed approach is evaluated in facial image reconstruction, facial super-resolution on two databases, i.e. LFPW and Helen. It is also evaluated on FG-NET database for the problem of age estimation.

****
Statistical Inference Models for Image Datasets With Systematic Variations
Won Hwa Kim, Barbara B. Bendlin, Moo K. Chung, Sterling C. Johnson, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4795-4803
Statistical analysis of longitudinal or cross sectionalbrain imaging data to identify effects of neurodegenerative diseases is a fundamental task in various studies in neuroscience. However, when there are systematic variations in the images due to parameters changes such as changes in the scanner protocol, hardware changes, or when combining data from multi-site studies, the statistical analysis becomes problematic. Motivated by this scenario, the goal of this paper is to develop a unified statistical solution to the problem of systematic variations in statistical image analysis. Based in part on recent literature in harmonic analysis on diffusion maps, we propose an algorithm which compares operators that are resilient to the systematic variations described above. These operators are derived from the empirical measurements of the image data and provide an efficient surrogate to capturing the actual changes across images. We also establish a connection between our method to the design of Wavelets in non-Euclidean space. To evaluate the proposed ideas, we present various experimental results on detecting changes in simulations as well as show how the method offers improved statistical power in the analysis of longitudinal real PIB-PET imaging data acquired from participants at risk for Alzheimer's disease(AD).
****
Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues
Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, Lubomir Bourdev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4804-4813
We explore the task of recognizing peoples' identities in photo albums in an unconstrained setting. To facilitate this, we introduce the new People In Photo Albums (PIPA) dataset, consisting of over 60000 instances of ~2000 individuals collected from public Flickr photo albums. With only about half of the person images containing a frontal face, the recognition task is very challenging due to the large variations in pose, clothing, camera viewpoint, image resolution and illumination. We propose the Pose Invariant PErson Recognition (PIPER) method, which accumulates the cues of poselet-level person recognizers trained by deep convolutional networks to discount for the pose variations, combined with a face recognizer and a global recognizer. Experiments on three different settings confirm that in our unconstrained setup PIPER significantly improves on the performance of DeepFace, which is one of the best face recognizers as measured on the LFW dataset.
****
Superpixel-Based Video Object Segmentation Using Perceptual Organization and Location Prior
Daniela Giordano, Francesca Murabito, Simone Palazzo, Concetto Spampinato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4814-4822
In this paper we present an approach for segmenting objects in videos taken in complex scenes with multiple and different targets. The method does not make any specific assumptions about the videos and relies on how objects are perceived by humans according to Gestalt laws. Initially, we rapidly generate a coarse foreground segmentation, which provides predictions about motion regions by analyzing how superpixel segmentation changes in consecutive frames. We then exploit these location priors to refine the initial segmentation by optimizing an energy function based on appearance and perceptual organization, only on regions where motion is observed. We evaluated our method on complex and challenging video sequences and it showed significant performance improvements over recent state-of-the-art methods, being also fast enough to be used for "on-the-fly" processing.
****
Robust Image Filtering Using Joint Static and Dynamic Guidance
Bumsub Ham, Minsu Cho, Jean Ponce; Proceedings of the IEEE Conference on Compute

r Vision and Pattern Recognition (CVPR), 2015, pp. 4823-4831
Regularizing images under a guidance signal has been used in various tasks in co mputer vision and computational photography, particularly for noise reduction an d joint upsampling. The aim is to transfer fine structures of guidance signals t o input images, restoring noisy or altered structures. One of main drawbacks in such a data-dependent framework is that it does not handle differences in struct ure between guidance and input images. We address this problem by jointly levera ging structural information of guidance and input images. Image filtering is for mulated as a nonconvex optimization problem, which is solved by the majorization -minimization algorithm. The proposed algorithm converges quickly while guarante eing a local mini- mum. It effectively controls image structures at different sc ales and can handle a variety of types of data from different sensors. We demons trate the flexibility and effectiveness of our model in several applications inc luding depth super- resolution, scale-space filtering, texture removal, flash/no n- flash denoising, and RGB/NIR denoising.
****

Solving Multiple Square Jigsaw Puzzles With Missing Pieces
Genady Paikin, Ayellet Tal; Proceedings of the IEEE Conference on Computer Visio n and Pattern Recognition (CVPR), 2015, pp. 4832-4839
Jigsaw-puzzle solving is necessary in many applications, including biology, arch aeology, and every-day life. In this paper we consider the square jigsaw puzzle problem, where the goal is to reconstruct the image from a set of non-overlappin g, unordered, square puzzle parts. Our key contribution is a fast, fully-automat ic, and general solver, which assumes no prior knowledge about the original imag e. It is general in the sense that it can handle puzzles of unknown size, with p ieces of unknown orientation, and even puzzles with missing pieces. Moreover, it can handle all the above, given pieces from multiple puzzles. Through an extens ive evaluation we show that our approach outperforms state-of-the-art methods on commonly-used datasets.
****

A Dynamic Convolutional Layer for Short Range Weather Prediction
Benjamin Klein, Lior Wolf, Yehuda Afek; Proceedings of the IEEE Conference on Co mputer Vision and Pattern Recognition (CVPR), 2015, pp. 4840-4848
We present a new deep network layer called ``Dynamic Convolutional Layer" which is a generalization of the convolutional layer. The conventional convolutional l ayer uses filters that are learned during training and are held constant during testing. In contrast, the dynamic convolutional layer uses filters that will var y from input to input during testing. This is achieved by learning a function th at maps the input to the filters. We apply the dynamic convolutional layer to th e application of short range weather prediction and show performance improvement s compared to other baselines.
****

SWIFT: Sparse Withdrawal of Inliers in a First Trial
Maryam Jaberi, Marianna Pensky, Hassan Foroosh; Proceedings of the IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4849-4857
We study the simultaneous detection of multiple structures in the presence of ov erwhelming number of outliers in a large population of points. Our approach redu ces the problem to sampling an extremely sparse subset of the original populatio n of data in one grab, followed by an unsupervised clustering of the population based on a set of instantiated models from this sparse subset. We show that the problem can be modeled using a multivariate hypergeometric distribution, and der ive accurate mathematical bounds to determine a tight approximation to the sampl e size, leading thus to a sparse sampling strategy. We evaluate the method thoro ughly in terms of accuracy, its behavior against varying input parameters, and c omparison against existing methods, including the state of the art. The key feat ures of the proposed approach are: (i) sparseness of the sampled set, where the level of sparseness is independent of the population size and the distribution o f data, (ii) robustness in the presence of overwhelming number of outliers, and (iii) unsupervised detection of all model instances, i.e. without requiring any prior knowledge of the number of embedded structures. To demonstrate the generic

nature of the proposed method, we show experimental results on different computer vision problems, such as detection of physical structures e.g. lines, planes, etc., as well as more abstract structures such as fundamental matrices, and homographies in multi-body structure from motion.
****

VIP: Finding Important People in Images
Clint Solomon Mathialagan, Andrew C. Gallagher, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4858-4866
People preserve memories of events such as birthdays, weddings, or vacations by capturing photos, often depicting groups of people. Invariably, some individuals in the image are more important than others given the context of the event. This paper analyzes the concept of the importance of individuals in group photographs. We address two specific questions - Given an image, who are the most important individuals in it? Given multiple images of a person, which image depicts the person in the most important role? We introduce a measure of importance of people in images and investigate the correlation between importance and visual saliency. We find that not only can we automatically predict the importance of people from purely visual cues, incorporating this predicted importance results in significant improvement in applications such as im2text (generating sentences that describe images of groups of people).
****

Dataset Fingerprints: Exploring Image Collections Through Data Mining
Konstantinos Rematas, Basura Fernando, Frank Dellaert, Tinne Tuytelaars; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4867-4875
As the amount of visual data increases, so does the need for summarization tools that can be used to explore large image collections and to quickly get familiar with their content. In this paper, we propose dataset fingerprints, a new and powerful method based on data mining that extracts meaningful patterns from a set of images. The discovered patterns are compositions of discriminative mid-level features that co-occur in several images. Compared to earlier work, ours stands out because i) it's fully unsupervised, ii) discovered patterns cover large parts of the images,often corresponding to full objects or meaningful parts thereof, and iii) different patterns are connected based on co-occurrence, allowing a user to ``browse'' / ``surf'' the images from one pattern to the next and to group patterns in a semantically meaningful manner.
****

Transport-Based Single Frame Super Resolution of Very Low Resolution Face Images
Soheil Kolouri, Gustavo K. Rohde; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4876-4884
Extracting high-resolution information from highly degraded facial images is an important problem with several applications in science and technology. Here we describe a single frame super resolution technique that uses a transport-based formulation of the problem. The method consists of a training and a testing phase. In the training phase, a nonlinear Lagrangian model of high-resolution facial appearance is constructed fully automatically. In the testing phase, the resolution of a degraded image is enhanced by finding the model parameters that best fit the given low resolution data. We test the approach on two face datasets, namely the extended Yale Face Database B and the AR face datasets, and compare it to state of the art methods. The proposed method outperforms existing solutions in problems related to enhancing images of very low resolution.
****

3D Reconstruction in the Presence of Glasses by Acoustic and Stereo Fusion
Mao Ye, Yu Zhang, Ruigang Yang, Dinesh Manocha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4885-4893
We present a practical and inexpensive method to reconstruct 3D scenes that include piece-wise planar transparent objects. Our work is motivated by the need for automatically generating 3D models of interior scenes, in which glass structures are common. These large structures are often invisible to cameras or even our

human visual system. Existing 3D reconstruction methods for transparent objects are usually not applicable in such a room-size reconstruction setting. Our approach augments a regular depth camera (e.g., the Microsoft Kinect camera) with a single ultrasonic sensor, which is able to measure distance to any objects, including transparent surfaces. We present a novel sensor fusion algorithm that first segments the depth map into different categories such as opaque/transparent/infinity (e.g., too far to measure) and then updates the depth map based on the segmentation outcome. Our current hardware setup can generate only one additional point measurement per frame, yet our fusion algorithm is able to generate satisfactory reconstruction results based on our probabilistic model. We highlight the performance in many challenging indoor benchmarks.
****

Deep Sparse Representation for Robust Image Registration
Yeqing Li, Chen Chen, Fei Yang, Junzhou Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4894-4901
The definition of the similarity measure is an essential component in image registration. In this paper, we propose a novel similarity measure for registration of two or more images. The proposed method is motivated by that the optimally registered images can be deeply sparsified in the gradient domain and frequency domain, with the separation of a sparse tensor of errors. One of the key advantages of the proposed similarity measure is its robustness to severe intensity distortions, which widely exist on medical images, remotely sensed images and natural photos due to the difference of acquisition modalities or illumination conditions. Two efficient algorithms are proposed to solve the batch image registration and pair registration problems in a unified framework. We validate our method on extensive challenging datasets. The experimental results demonstrate the robustness, accuracy and efficiency of our method over 9 traditional and state-of-the-art algorithms on synthetic images and a wide range of real-world applications.
****

Real-Time Part-Based Visual Tracking via Adaptive Correlation Filters
Ting Liu, Gang Wang, Qingxiong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4902-4912
Robust object tracking is a challenging task in computer vision. To better solve the partial occlusion issue, part-based methods are widely used in visual object trackers. However, due to the complicated online training and updating process, most of these part-based trackers cannot run in real-time. Correlation filters have been used in tracking tasks recently because of the high efficiency. However, the conventional correlation filter based trackers cannot deal with occlusion. Furthermore, most correlation filter based trackers fix the scale and rotation of the target which makes the trackers unreliable in long-term tracking tasks. In this paper, we propose a novel tracking method which track objects based on parts with multiple correlation filters. Our method can run in real-time. Additionally, the Bayesian inference framework and a structural constraint mask are adopted to enable our tracker to be robust to various appearance changes. Extensive experiments have been done to prove the effectiveness of our method.
****

Beyond Spatial Pooling: Fine-Grained Representation Learning in Multiple Domains
Chi Li, Austin Reiter, Gregory D. Hager; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4913-4922
Object recognition systems have shown great progress over recent years. However, creating object representations that are robust to changes in viewpoint while capturing local visual details continues to be a challenge. In particular, recent convolutional architectures employ spatial pooling to achieve scale and shift invariances, but they are still sensitive to out-of-plane rotations. In this paper, we formulate a probabilistic framework for analyzing the performance of pooling. This framework suggests two directions for improvement. First, we apply multiple scales of filters coupled with different pooling granularities, and second we make use of color as an additional pooling domain, thereby reducing the sensitivity to spatial deformations. We evaluate our algorithm on the object instance recognition task using two independent publicly available RGB-D datasets, and d

emonstrate significant improvements over the current state-of-the-art. In additi
on, we present a new dataset for industrial objects to further validate the effe
ctiveness of our approach versus other state-of-the-art approaches for object re
cognition using RGB-D data.
****

HC-Search for Structured Prediction in Computer Vision
Michael Lam, Janardhan Rao Doppa, Sinisa Todorovic, Thomas G. Dietterich; Procee
dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
2015, pp. 4923-4932
The mainstream approach to structured prediction problems in computer vision is
to learn an energy function such that the solution minimizes that function. At
prediction time, this approach must solve an often-challenging optimization prob
lem. Search-based methods provide an alternative that has the potential to achi
eve higher performance. These methods learn to control a search procedure that
constructs and evaluates candidate solutions. The recently-developed HC-Search
method has been shown to achieve state-of-the-art results in natural language pr
ocessing, but mixed success when applied to vision problems. This paper studies
whether HC-Search can achieve similarly competitive performance on basic vision
tasks such as object detection, scene labeling, and monocular depth estimation,
where the leading paradigm is energy minimization. To this end, we introduce a
search operator suited to the vision domain that improves a candidate solution
by probabilistically sampling likely object configurations in the scene from the
hierarchical Berkeley segmentation. We complement this search operator by appl
ying the DAgger algorithm to robustly train the search heuristic so it learns fr
om its previous mistakes. Our evaluation shows that these improvements reduce t
he branching factor and search depth, and thus give a significant performance bo
ost. Our state-of-the-art results on scene labeling and depth estimation sugges
t that HC-Search provides a suitable tool for learning and inference in vision.
****

Revisiting Kernelized Locality-Sensitive Hashing for Improved Large-Scale Image
Retrieval
Ke Jiang, Qichao Que, Brian Kulis; Proceedings of the IEEE Conference on Compute
r Vision and Pattern Recognition (CVPR), 2015, pp. 4933-4941
We present a simple but powerful reinterpretation of kernelized locality-sensiti
ve hashing (KLSH), a general and popular method developed in the vision communit
y for performing approximate nearest-neighbor searches in an arbitrary reproduci
ng kernel Hilbert space (RKHS). Our new perspective is based on viewing the ste
ps of the KLSH algorithm in an appropriately projected space, and has several ke
y theoretical and practical benefits. First, it eliminates the problematic conc
eptual difficulties that are present in the existing motivation of KLSH. Second
, it yields the first formal retrieval performance bounds for KLSH. Third, our
analysis reveals two techniques for boosting the empirical performance of KLSH.
We evaluate these extensions on several large-scale benchmark image retrieval d
ata sets, and show that our analysis leads to improved recall performance of at
least 12%, and sometimes much higher, over the standard KLSH method.
****

High-Speed Hyperspectral Video Acquisition With a Dual-Camera Architecture
Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, Wenjun Zeng, Feng Wu; Procee
dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
2015, pp. 4942-4950
We propose a novel dual-camera design to acquire 4D high-speed hyperspectral (HS
HS) videos with high spatial and spectral resolution. Our work has two key techn
ical contributions. First, we build a dual-camera system that simultaneously cap
tures a panchromatic video at a high frame rate and a hyperspectral video at a l
ow frame rate, which jointly provide reliable projections for the underlying HSH
S video. Second, we exploit the panchromatic video to learn an over-complete 3D
dictionary to represent each band-wise video sparsely, and a robust computationa
l reconstruction is then employed to recover the HSHS video based on the joint v
ideos and the self-learned dictionary. Experimental results demonstrate that, fo
r the first time to our knowledge, the hyperspectral video frame rate reaches up

to 100fps with decent quality, even when the incident light is not strong.
****

More About VLAD: A Leap From Euclidean to Riemannian Manifolds
Masoud Faraki, Mehrtash T. Harandi, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4951-4960
This paper takes a step forward in image and video coding by extending the well-known Vector of Locally Aggregated Descriptors (VLAD) onto an extensive space of curved Riemannian manifolds. We provide a comprehensive mathematical framework that formulates the aggregation problem of such manifold data into an elegant solution. In particular, we consider structured descriptors from visual data, namely Region Covariance Descriptors and linear subspaces that reside on the manifold of Symmetric Positive Definite matrices and the Grassmannian manifolds, respectively. Through rigorous experimental validation, we demonstrate the superior performance of this novel Riemannian VLAD descriptor on several visual classification tasks including video-based face recognition, dynamic scene recognition, and head pose classification.
****

Camera Intrinsic Blur Kernel Estimation: A Reliable Framework
Ali Mosleh, Paul Green, Emmanuel Onzon, Isabelle Begin, J.M. Pierre Langlois; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4961-4968
This paper presents a reliable non-blind method to measure intrinsic lens blur. We first introduce an accurate camera-scene alignment framework that avoids erroneous homography estimation and camera tone curve estimation. This alignment is used to generate a sharp correspondence of a target pattern captured by the camera. Second, we introduce a Point Spread Function (PSF) estimation approach where information about the frequency spectrum of the target image is taken into account. As a result of these steps and the ability to use multiple target images in this framework, we achieve a PSF estimation method robust against noise and suitable for mobile devices. Experimental results show that the proposed method results in PSFs with more than 10 dB higher accuracy in noisy conditions compared with the PSFs generated using state-of-the-art techniques.
****

Classifier Learning With Hidden Information
Ziheng Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4969-4977
Traditional data-driven classifier learning approaches become limited when the training data is inadequate either in quantity or quality. To address this issue, in this paper we propose to combine hidden information and data to enhance classifier learning. Hidden information represents information that is only available during training but not available during testing. It often exists in many applications yet has not been thoroughly exploited, and existing methods to utilize hidden information are still limited. To this end, we propose two general approaches to exploit different types of hidden information to improve different classifiers. We also extend the proposed methods to deal with incomplete hidden information. Experimental results on different applications demonstrate the effectiveness of the proposed methods for exploiting hidden information and their superior performance to existing methods.
****

Single Target Tracking Using Adaptive Clustered Decision Trees and Dynamic Multi-Level Appearance Models
Jingjing Xiao, Rustam Stolkin, Ales Leonardis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4978-4987
This paper presents a method for single target tracking of arbitrary objects in challenging video sequences. Targets are modeled at three different levels of granularity (pixel level, parts-based level and bounding box level), which are cross-constrained to enable robust model relearning. The main contribution is an adaptive clustered decision tree method which dynamically selects the minimum combination of features necessary to sufficiently represent each target part at each frame, thereby providing robustness with computational efficiency. The adaptive

clustered decision tree is implemented in two separate parts of the tracking algorithm: firstly to enable robust matching at the parts-based level between successive frames; and secondly to select the best superpixels for learning new parts of the target. We have tested the tracker using two different tracking benchmarks (VOT2013-2014 and CVPR2013 tracking challenges), based on two different test methodologies, and show it to be significantly more robust than the best state-of-the-art methods from both of those tracking challenges, while also offering competitive tracking precision.
****

Simultaneous Video Defogging and Stereo Reconstruction
Zhuwen Li, Ping Tan, Robby T. Tan, Danping Zou, Steven Zhiying Zhou, Loong-Fah Cheong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4988-4997
We present a method to jointly estimate scene depth and recover the clear latent image from a foggy video sequence. In our formulation, the depth cues from stereo matching and fog information reinforce each other, and produce superior results than conventional stereo or defogging algorithms. We first improve the photo-consistency term to explicitly model the appearance change due to the scattering effects. The prior matting Laplacian constraint on fog transmission imposes a detail-preserving smoothness constraint on the scene depth. We further enforce the ordering consistency between scene depth and fog transmission at neighboring points. These novel constraints are formulated together in an MRF framework, which is optimized iteratively by introducing auxiliary variables. The experiment results on real videos demonstrate the strength of our method.
****

Face Alignment by Coarse-to-Fine Shape Searching
Shizhan Zhu, Cheng Li, Chen Change Loy, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4998-5006
We present a novel face alignment framework based on coarse-to-fine shape searching. Unlike the conventional cascaded regression approaches that start with an initial shape and refine the shape in a cascaded manner, our approach begins with a coarse search over a shape space that contains diverse shapes, and employs the coarse solution to constrain subsequent finer search of shapes. The unique stage-by-stage progressive and adaptive search i) prevents the final solution from being trapped in local optima due to poor initialisation, a common problem encountered by cascaded regression approaches; and ii) improves the robustness in coping with large pose variations. The framework demonstrates real-time performance and state-of-theart results on various benchmarks including the challenging 300-W dataset.
****

Learning Deep Representations for Ground-to-Aerial Geolocalization
Tsung-Yi Lin, Yin Cui, Serge Belongie, James Hays; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5007-5015
The recent availability of geo-tagged images and rich geospatial data has inspired a number of algorithms for image based geolocalization. Most approaches predict the location of a query image by matching to ground-level images with known locations (e.g., street-view data). However, most of the Earth does not have ground-level reference photos available. Fortunately, more complete coverage is provided by oblique aerial or "bird's eye" imagery. In this work, we localize a ground-level query image by matching it to a reference database of aerial imagery. We use publicly available data to build a dataset of 78K aligned cross- view image pairs. The primary challenge for this task is that traditional computer vision approaches cannot handle the wide baseline and appearance variation of these cross- view pairs. We use our dataset to learn a feature representation in which matching views are near one another and mismatched views are far apart. Our proposed approach, Where-CNN, is inspired by deep learning success in face verification and achieves significant improvements over traditional hand-crafted features and existing deep features learned from other large-scale databases. We show the effectiveness of Where-CNN in finding matches between street view and aerial view imagery and demonstrate the ability of our learned features to generalize to

novel locations.
****
Unsupervised Simultaneous Orthogonal Basis Clustering Feature Selection
Dongyoon Han, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5016-5023
In this paper, we propose a novel unsupervised feature selection method: Simultaneous Orthogonal basis Clustering Feature Selection (SOCFS). To perform feature selection on unlabeled data effectively, a regularized regression-based formulation with a new type of target matrix is designed. The target matrix captures latent cluster centers of the projected data points by performing orthogonal basis clustering, and then guides the projection matrix to select discriminative features. Unlike the recent unsupervised feature selection methods, SOCFS does not explicitly use the pre-computed local structure information for data points represented as additional terms of their objective functions, but directly computes latent cluster information by the target matrix conducting orthogonal basis clustering in a single unified term of the proposed objective function. It turns out that the proposed objective function can be minimized by a simple optimization algorithm. Experimental results demonstrate the effectiveness of SOCFS achieving the state-of-the-art results with diverse real world datasets.
****
Space-Time Tree Ensemble for Action Recognition
Shugao Ma, Leonid Sigal, Stan Sclaroff; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5024-5032
Human actions are, inherently, structured patterns of body movements. We explore ensembles of hierarchical spatio-temporal trees, discovered directly from training data, to model these structures for action recognition. The hierarchical spatio-temporal trees provide a robust mid-level representation for actions. However, discovery of frequent and discriminative tree structures is challenging due to the exponential search space, particularly if one allows partial matching. We address this by first building a concise action vocabulary via discriminative clustering. Using the action vocabulary we then utilize tree mining with subsequent tree clustering and ranking to select a compact set of highly discriminative tree patterns. We show that these tree patterns, alone, or in combination with shorter patterns (action words and pair-wise patterns) achieve state-of-the-art performance on two challenging datasets: UCF Sports and HighFive. Moreover, trees learned on HighFive are used in recognizing two action classes in a different dataset, Hollywood3D, demonstrating the potential for cross-dataset generality of the trees our approach discovers.
****
Subgraph Decomposition for Multi-Target Tracking
Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5033-5041
Tracking multiple targets in a video, based on a finite set of detection hypotheses, is a persistent problem in computer vision. A common strategy for tracking is to first select hypotheses spatially and then to link these over time while maintaining disjoint path constraints. In crowded scenes multiple hypotheses will often be similar to each other making selection of optimal links an unnecessary hard optimization problem due to the sequential treatment of space and time. Embracing this observation, we propose to link and cluster plausible detections jointly across space and time. Specifically, we state multi-target tracking as a Minimum Cost Subgraph Multicut Problem. Evidence about pairs of detection hypotheses is incorporated whether the detections are in the same frame, neighboring frames or distant frames. This facilitates long-range re-identification and within-frame clustering. Results for published benchmark sequences demonstrate the superiority of this approach.
****
Understanding Image Structure via Hierarchical Shape Parsing
Xian-Ming Liu, Rongrong Ji, Changhu Wang, Wei Liu, Bineng Zhong, Thomas S. Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), 2015, pp. 5042-5050
Exploring image structure is a long-standing yet important research subject in t
he computer vision community. In this paper, we focus on understanding image str
ucture inspired by the "simple-to-complex" biological evidence. A hierarchical s
hape parsing strategy is proposed to partition and organize image components int
o a hierarchical structure in the scale space. To improve the robustness and fle
xibility of image representation, we further bundle the image appearances into h
ierarchical parsing trees. Image descriptions are subsequently constructed by pe
rforming a structural pooling, facilitating efficient matching between the parsi
ng trees. We leverage the proposed hierarchical shape parsing to study two exemp
lar applications including edge scale refinement and unsupervised "objectness" d
etection. We show competitive parsing performance comparing to the state-of-the-
arts in above scenarios with far less proposals, which thus demonstrates the adv
antage of the proposed parsing scheme.
****

Coarse-To-Fine Region Selection and Matching
Yanchao Yang, Zhaojin Lu, Ganesh Sundaramoorthi; Proceedings of the IEEE Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5051-5059
We present a new approach to wide baseline matching. We propose to use a hierarc
hical decomposition of the image domain and coarse-to-fine selection of regions
to match. In contrast to interest point matching methods, which sample salient r
egions to reduce the cost of comparing all regions in two images, our method eli
minates regions systematically to achieve efficiency. One advantage of our appro
ach is that it is not restricted to covariant salient regions, which is too rest
rictive under large viewpoint and leads to few corresponding regions. Affine inv
ariant matching of regions in the hierarchy is achieved efficiently by a coarse-
to-fine search of the affine space. Experiments on two benchmark datasets shows
that our method finds more correct correspondence of the image (with fewer false
 alarms) than other wide baseline methods on large viewpoint change.
****

Label Consistent Quadratic Surrogate Model for Visual Saliency Prediction
Yan Luo, Yongkang Wong, Qi Zhao; Proceedings of the IEEE Conference on Computer
Vision and Pattern Recognition (CVPR), 2015, pp. 5060-5069
Recently, an increasing number of works have proposed to learn visual saliency b
y leveraging human fixations. However, the collection of human fixations is time
 consuming and the existing eye tracking datasets are generally small when compa
red with other domains. Thus, it contains a certain degree of dataset bias due t
o the large image variations (e.g., outdoor scenes vs. emotion-evoking images).
In the learning based saliency prediction literature, most models are trained an
d evaluated within the same dataset and cross dataset validation is not yet a co
mmon practice. Instead of directly applying model learned from another dataset i
n cross dataset fashion, it is better to transfer the prior knowledge obtained f
rom one dataset to improve the training and prediction on another. In addition,
since new datasets are built and shared in the community from time to time, it w
ould be good not to retrain the entire model when new data are added. To address
 these problems, we proposed a new learning based saliency model, namely Label C
onsistent Quadratic Surrogate algorithm, which employs an iterative online algor
ithm to learn a sparse dictionary with label consistent constraint. The advantag
es of the proposed model are three-folds: (1) the quadratic surrogate function g
uarantees convergence at each iteration, (2) the label consistent constraint enf
orces the predicted sparse code to be discriminative, and (3) the online propert
ies enable the proposed algorithm to adapt existing model with new data without
retraining. As shown in this work, the proposed saliency model achieves better p
erformance than the state-of-the-art saliency models.
****

Subgraph Matching Using Compactness Prior for Robust Feature Correspondence
Yumin Suh, Kamil Adamczewski, Kyoung Mu Lee; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5070-5078
Feature correspondence plays a central role in various computer vision applicati
ons. It is widely formulated as a graph matching problem due to its robust perfo

rmance under challenging conditions, such as background clutter, object deformation and repetitive patterns. A variety of fast and accurate algorithms have been proposed for graph matching. However, most of them focus on improving the recall of the solution while rarely considering its precision, thus inducing a solution with numerous outliers. To address both precision and recall feature correspondence should rather be formulated as a subgraph matching problem. This paper proposes a new subgraph matching formulation which uses a compactness prior, an additional constraint that prefers sparser solutions and effectively eliminates outliers. To solve the new optimization problem, we propose a meta-algorithm based on Markov chain Monte Carlo. By constructing Markov chain on the restricted search space instead of the original solution space, our method approximates the solution effectively. The experiments indicate that our proposed formulation and algorithm significantly improve the baseline performance under challenging conditions when both outliers and deformation noise are present.

****

Pedestrian Detection Aided by Deep Learning Semantic Tasks

Yonglong Tian, Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5079-5087

Deep learning methods have achieved great success in pedestrian detection, owing to its ability to learn features from raw pixels. However, they can mainly capture middlelevel representations, such as pose of pedestrian, but confuses positive with hard negative samples (Fig.1 (a)), which have large ambiguity and can only be distinguished by highlevel representation. To address this ambiguity, this work jointly optimize pedestrian detection with semantic tasks, including pedestrian attributes (e.g. 'carrying backpack') and scene attributes (e.g. 'vehicle', 'tree', and 'horizontal'). Rather than expensively annotating scene attributes, we transfer attributes information from existing scene segmentation datasets to the pedestrian dataset, by proposing a novel deep model to learn high-level features from multiple tasks and multiple data sources. Since distinct tasks have distinct convergence rates and data from different datasets have different distributions, a multi-task objective function is carefully designed to coordinate tasks and reduce discrepancies among datasets. The importance coefficients of tasks and network parameters in this objective function can be iteratively estimated. Extensive evaluations show that the proposed approach outperforms the state-of-the-art on the challenging Caltech [10] and ETH [11] datasets where it reduces the miss rates of previous deep models by 17 and 5.5 percent, respectively.

****

Multihypothesis Trajectory Analysis for Robust Visual Tracking

Dae-Youn Lee, Jae-Young Sim, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5088-5096

The notion of multihypothesis trajectory analysis (MTA) for robust visual tracking is proposed in this work. We employ multiple component trackers using texture, color, and illumination invariant features, respectively. Each component tracker traces a target object forwardly and then backwardly over a time interval. By analyzing the pair of the forward and backward trajectories, we measure the robustness of the component tracker. To this end, we extract the geometry similarity, the cyclic weight, and the appearance similarity from the forward and backward trajectories. We select the optimal component tracker to yield the maximum robustness score, and use its forward trajectory as the final tracking result. Experimental results show that the proposed MTA tracker improves the robustness and the accuracy of tracking, outperforming the state-of-the-art trackers on a recent benchmark dataset.

****

Domain-Size Pooling in Local Descriptors: DSP-SIFT

Jingming Dong, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5097-5106

We introduce a simple modification of local image descriptors, such as SIFT, based on pooling gradient orientations across different domain sizes, in addition to spatial locations. The resulting descriptor, which we call DSP-SIFT, outperforms other methods in wide-baseline matching benchmarks, including those based on

convolutional neural networks, despite having the same dimension of SIFT and req
uiring no training.
****

Object Detection by Labeling Superpixels
Junjie Yan, Yinan Yu, Xiangyu Zhu, Zhen Lei, Stan Z. Li; Proceedings of the IEEE
 Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5107-51
16
Object detection is always conducted by object proposal generation and classific
ation sequentially. This paper handles object detection in a superpixel oriented
 manner instead of the proposal oriented. Specially, this paper takes object det
ection as a multi-label superpixel labeling problem by minimizing an energy func
tion. It uses the data cost term to capture the appearance, smooth cost term to
encode the spatial context and label cost term to favor compact detection. The d
ata cost is learned through a convolutional neural network and the parameters in
 the labeling model are learned through a structural SVM. Compared with proposal
 generation and classification based methods, the proposed superpixel labeling m
ethod can naturally detect objects missed by proposal generation step and captur
e the global image context to infer the overlapping objects. The proposed meth
od shows its advantage in Pascal VOC and ImageNet. Notably, it performs better t
han the ImageNet ILSVRC2014 winner GoogLeNet (45.0% V.S. 43.9% in mAP) with much
 shallower and fewer CNNs.
****

Fast 2D Border Ownership Assignment
Ching Teo, Cornelia Fermuller, Yiannis Aloimonos; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5117-5125
A method for efficient border ownership assignment in 2D images is proposed. Lev
eraging on recent advances using Structured Random Forests (SRF) for boundary de
tection, we impose a novel border ownership structure that detects both boundari
es and border ownership at the same time. Key to this work are features that pre
dict ownership cues from 2D images. To this end, we use several different local
cues: shape, spectral properties of boundary patches, and semi-global grouping c
ues that are indicative of perceived depth. For shape, we use HoG-like descripto
rs that encode local curvature (convexity and concavity). For spectral propertie
s, such as extremal edges, we first learn an orthonormal basis spanned by the to
p K eigenvectors via PCA over common types of contour tokens. For grouping, we i
ntroduce a novel mid-level descriptor that captures patterns near edges and indi
cates ownership information of the boundary. Experimental results over a subset
of the Berkeley Segmentation Dataset (BSDS) and the NYU Depth V2 dataset show th
at our method's performance exceeds current state-of-the-art multi-stage approac
hes that use more complex features.
****

From Single Image Query to Detailed 3D Reconstruction
Johannes L. Schonberger, Filip Radenovic, Ondrej Chum, Jan-Michael Frahm; Procee
dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
2015, pp. 5126-5134
Structure-from-Motion for unordered image collections has significantly advanced
 in scale over the last decade. This impressive progress can be in part attribut
ed to the introduction of efficient retrieval methods for those systems. While t
his boosts scalability, it also limits the amount of detail that the large-scale
 reconstruction systems are able to produce. In this paper, we propose a joint r
econstruction and retrieval system that maintains the scalability of large-scale
 Structure-from-Motion systems while also recovering the often lost ability of r
econstructing fine details of the scene. We demonstrate our proposed method on a
 large-scale dataset of 7.4 million images downloaded from the Internet.
****

Fast and Flexible Convolutional Sparse Coding
Felix Heide, Wolfgang Heidrich, Gordon Wetzstein; Proceedings of the IEEE Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5135-5143
Convolutional sparse coding (CSC) has become an increasingly important tool in m
achine learning and computer vision. Image features can be learned and subsequen

tly used for classification and reconstruction tasks. As opposed to patch-based methods, convolutional sparse coding operates on whole images, thereby seamlessly capturing the correlation between local neighborhoods. In this paper, we propose a new approach to solving CSC problems and show that our method converges significantly faster and also finds better solutions than the state of the art. In addition, the proposed method is the first efficient approach to allow for proper boundary conditions to be imposed and it also supports feature learning from incomplete data as well as general reconstruction problems.
****

Iteratively Reweighted Graph Cut for Multi-Label MRFs With Non-Convex Priors
Thalaiyasingam Ajanthan, Richard Hartley, Mathieu Salzmann, Hongdong Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5144-5152
While widely acknowledged as highly effective in computer vision, multi-label MRFs with non-convex priors are difficult to optimize. To tackle this, we introduce an algorithm that iteratively approximates the original energy with an appropriately weighted surrogate energy that is easier to minimize. Our algorithm guarantees that the original energy decreases at each iteration. In particular, we consider the scenario where the global minimizer of the weighted surrogate energy can be obtained by a multi-label graph cut algorithm, and show that our algorithm then lets us handle of large variety of non-convex priors. We demonstrate the benefits of our method over state-of-the-art MRF energy minimization techniques on stereo and inpainting problems.
****

Pairwise Geometric Matching for Large-Scale Object Retrieval
Xinchao Li, Martha Larson, Alan Hanjalic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5153-5161
Spatial verification is a key step in boosting the performance of object-based image retrieval. It serves to eliminate unreliable correspondences between salient points in a given pair of images and is typically performed by analyzing the consistency of spatial transformations between the image regions involved in individual correspondences. In this paper, we consider the pairwise geometric relations between correspondences and propose a strategy to incorporate these relations at significantly reduced computational cost, which makes it suitable for large-scale object retrieval. In addition, we combine the information on geometric relations from both the individual correspondences and pairs of correspondences to further improve the verification accuracy. Experimental results on three reference datasets show that the proposed approach results in a substantial performance improvement compared to the existing methods, without making concessions regarding computational efficiency.
****

Deep Convolutional Neural Fields for Depth Estimation From a Single Image
Fayao Liu, Chunhua Shen, Guosheng Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5162-5170
We consider the problem of depth estimation from a single monocular image in this work. It is a challenging task as no reliable depth cues are available, e.g., stereo correspondences, motions etc. Previous efforts have been focusing on exploiting geometric priors or additional sources of information, with all using hand-crafted features. Recently, there is mounting evidence that features from deep convolutional neural networks (CNN) are setting new records for various vision applications. On the other hand, considering the continuous characteristic of the depth values, depth estimations can be naturally formulated into a continuous conditional random field (CRF) learning problem. Therefore, we in this paper present a deep convolutional neural field model for estimating depths from a single image, aiming to jointly explore the capacity of deep CNN and continuous CRF. Specifically, we propose a deep structured learning scheme which learns the unary and pairwise potentials of continuous CRF in a unified deep CNN framework. The proposed method can be used for depth estimations of general scenes with no geometric priors nor any extra information injected. In our case, the integral of the partition function can be analytically calculated, thus we can exactly solve t

he log-likelihood optimization. Moreover, solving the MAP problem for predicting depths of a new image is highly efficient as closed-form solutions exist. We experimentally demonstrate that the proposed method outperforms state-of-the-art depth estimation methods on both indoor and outdoor scene datasets.
****

Data-Driven Sparsity-Based Restoration of JPEG-Compressed Images in Dual Transform-Pixel Domain
Xianming Liu, Xiaolin Wu, Jiantao Zhou, Debin Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5171-5178
Arguably the most common cause of image degradation is compression. This papers presents a novel approach to restoring JPEG-compressed images. The main innovation is in the approach of exploiting residual redundancies of JPEG code streams and sparsity properties of latent images. The restoration is a sparse coding process carried out jointly in the DCT and pixel domains. The prowess of the proposed approach is directly restoring DCT coefficients of the latent image to prevent the spreading of quantization errors into the pixel domain, and at the same time using on-line machine-learnt local spatial features to regulate the solution of the underlying inverse problem. Experimental results are encouraging and show the promise of the new approach in significantly improving the quality of DCT-coded images.
****

TVSum: Summarizing Web Videos Using Titles
Yale Song, Jordi Vallmitjana, Amanda Stent, Alejandro Jaimes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5179-5187
Video summarization is a challenging problem in part because knowing which part of a video is important requires prior knowledge about its main topic. We present TVSum, an unsupervised video summarization framework that uses title-based image search results to find visually important shots. We observe that a video title is often carefully chosen to be maximally descriptive of its main topic, and hence images related to the title can serve as a proxy for important visual concepts of the main topic. However, because titles are free-formed, unconstrained, and often written ambiguously, images searched using the title can contain noise (images irrelevant to video content) and variance (images of different topics). To deal with this challenge, we developed a novel co-archetypal analysis technique that learns canonical visual concepts shared between video and images, but not in either alone, by finding a joint-factorial representation of two data sets. We introduce a new benchmark dataset, TVSum50, that contains 50 videos and their shot-level importance scores annotated via crowdsourcing. Experimental results on two datasets, SumMe and TVSum50, suggest our approach produces superior quality summaries compared to several recently proposed approaches.
****

Understanding Deep Image Representations by Inverting Them
Aravindh Mahendran, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5188-5196
Image representations, from SIFT and Bag of Visual Words to Convolutional Neural Networks (CNNs), are a crucial component of almost any image understanding system. Nevertheless, our understanding of them remains limited. In this paper we conduct a direct analysis of the visual information contained in representations by asking the following question: given an encoding of an image, to which extent is it possible to reconstruct the image itself? To answer this question we contribute a general framework to invert representations. We show that this method can invert representations such as HOG more accurately than recent alternatives while being applicable to CNNs too. We then use this technique to study the inverse of recent state-of-the-art CNN image representations for the first time. Among our findings, we show that several layers in CNNs retain photographically accurate information about the image, with different degrees of geometric and photometric invariance.
****

Single Image Super-Resolution From Transformed Self-Exemplars

Jia-Bin Huang, Abhishek Singh, Narendra Ahuja; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5197-5206
Self-similarity based super-resolution (SR) algorithms are able to produce visua lly pleasing results without extensive training on external databases. Such algo rithms exploit the statistical prior that patches in a natural image tend to rec ur within and across scales of the same image. However, the internal dictionary obtained from the given image may not always be sufficiently expressive to cover the textural appearance variations in the scene. In this paper, we extend self- similarity based SR to overcome this drawback. We expand the internal patch sear ch space by allowing geometric variations. We do so by explicitly localizing pla nes in the scene and using the detected perspective geometry to guide the patch search process. We also incorporate additional affine transformations to accommo date local shape variations. We propose a compositional model to simultaneously handle both types of transformations. We extensively evaluate the performance in both urban and natural scenes. Even without using any external training databas es, we achieve significantly superior results on urban scenes, while maintaining comparable performance on natural scenes as other state-of-the-art SR algorithm s.

****

Constrained Planar Cuts - Object Partitioning for Point Clouds
Markus Schoeler, Jeremie Papon, Florentin Worgotter; Proceedings of the IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5207-5215
While humans can easily separate unknown objects into meaningful parts, recent s egmentation methods can only achieve similar partitionings by training on human- annotated ground-truth data. Here we introduce a bottom-up method for segmenting 3D point clouds into functional parts which does not require supervision and ac hieves equally good results. Our method uses local concavities as an indicator f or inter-part boundaries. We show that this criterion is efficient to compute an d generalizes well across different object classes. The algorithm employs a nove l locally constrained geometrical boundary model which proposes greedy cuts thro ugh a local concavity graph. Only planar cuts are considered and evaluated using a cost function, which rewards cuts orthogonal to concave edges. Additionally, a local clustering constraint is applied to ensure the partitioning only affects relevant locally concave regions. We evaluate our algorithm on recordings from an RGB-D camera as well as the Princeton Segmentation Benchmark, using a fixed s et of parameters across all object classes. This stands in stark contrast to mos t reported results which require either knowing the number of parts or annotated ground-truth for learning. Our approach outperforms all existing bottom-up meth ods (reducing the gap to human performance by up to 50%) and achieves scores sim ilar to top-down data-driven approaches.

****

A Weighted Sparse Coding Framework for Saliency Detection
Nianyi Li, Bilin Sun, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5216-5223
There is an emerging interest on using high-dimensional datasets beyond 2D image s in saliency detection. Examples include 3D data based on stereo matching and K inect sensors and more recently 4D light field data. However, these techniques a dopt very different solution frameworks, in both type of features and procedures on using them. In this paper, we present a unified saliency detection framework for handling heterogenous types of input data. Our approach builds dictionaries using data-specific features. Specifically, we first select a group of potentia l background superpixels to build a primitive non-saliency dictionary. We then p rune the outliers in the dictionary and test on the remaining superpixels to ite ratively refine the dictionary. Comprehensive experiments show that our approach universally outperforms the state-of-the-art solution on all 2D, 3D and 4D data .

****

Handling Motion Blur in Multi-Frame Super-Resolution
Ziyang Ma, Renjie Liao, Xin Tao, Li Xu, Jiaya Jia, Enhua Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 522

4-5232
Ubiquitous motion blur easily fails multi-frame super-resolution (MFSR). Our method proposed in this paper tackles this issue by optimally searching least blurred pixels in MFSR. An EM framework is proposed to guide residual blur estimation and high-resolution image reconstruction. To suppress noise, we employ a family of sparse penalties as natural image priors, along with an effective solver. Theoretical analysis is performed on how and when our method works. The relationship between estimation errors of motion blur and the quality of input images is discussed. Our method produces sharp and higher-resolution results given input of challenging low-resolution noisy and blurred sequences.
****

Approximate Nearest Neighbor Fields in Video
Nir Ben-Zrihem, Lihi Zelnik-Manor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5233-5242
We introduce RIANN (Ring Intersection Approximate Nearest Neighbor search), an algorithm for matching patches of a video to a set of reference patches in real-time. For each query, RIANN finds potential matches by intersecting rings around key points in appearance space. Its search complexity is reversely correlated to the amount of temporal change, making it a good fit for videos, where typically most patches change slowly with time. Experiments show that RIANN is up to two orders of magnitude faster than previous ANN methods, and is the only solution that operates in real-time. We further demonstrate how RIANN can be used for real-time video processing and provide examples for a range of real-time video applications, including colorization, denoising, and several artistic effects.
****

Inverting RANSAC: Global Model Detection via Inlier Rate Estimation
Roee Litman, Simon Korman, Alexander Bronstein, Shai Avidan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5243-5251
This work presents a novel approach for detecting inliers in a given set of correspondences (matches). It does so without explicitly identifying any consensus set, based on a method for inlier rate estimation (IRE). Given such an estimator for the inlier rate, we also present an algorithm that detects a globally optimal transformation. We provide a theoretical analysis of the IRE method using a stochastic generative model on the continuous spaces of matches and transformations. This model allows rigorous investigation of the limits of our IRE method for the case of 2D-translation, further giving bounds and insights for the more general case. Our theoretical analysis is validated empirically and is shown to hold in practice for the more general case of 2D-affinities. In addition, we show that the combined framework works on challenging cases of 2D-homography estimation, with very few and possibly noisy inliers, where RANSAC generally fails.
****

Robust Multi-Image Based Blind Face Hallucination
Yonggang Jin, Christos-Savvas Bouganis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5252-5260
This paper proposes a robust multi-image based blind face hallucination framework to super-resolve LR faces. The proposed framework first estimates both blurring kernel and transformations of multiple LR faces by robust deblurring and registration in PCA subspace. A patch-wise mixture of probabilistic PCA prior is then incorporated for face super-resolution. Previous work on face SR using PCA prior can be viewed as special cases of the framework. Experimental results in both simulated and real LR sequences demonstrate very promising performance of the proposed method.
****

On Learning Optimized Reaction Diffusion Processes for Effective Image Restoration
Yunjin Chen, Wei Yu, Thomas Pock; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5261-5269
For several decades, image restoration remains an active research topic in low-level computer vision and hence new approaches are constantly emerging. However,

many recently proposed algorithms achieve state-of-the-art performance only at t
he expense of very high computation time, which clearly limits their practical r
elevance. In this work, we propose an effective approach with both high computat
ional efficiency and high restoration quality. We extend conventional nonlinear
reaction diffusion models by several parametrized linear filters as well as seve
ral parametrized influence functions. We propose to train the parameters of the
 filters and the influence functions through a loss based approach. Experiments
show that our trained nonlinear reaction diffusion models largely benefit from t
he training of the parameters and finally lead to the best reported performance
on common test datasets for image restoration. Due to their structural simplicit
y, our trained models are highly efficient and are also well-suited for parallel
 computation on GPUs.
****

A Flexible Tensor Block Coordinate Ascent Scheme for Hypergraph Matching
Quynh Nguyen, Antoine Gautier, Matthias Hein; Proceedings of the IEEE Conference
 on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5270-5278
The estimation of correspondences between two images resp. point sets is a core
problem in computer vision. One way to formulate the problem is graph matching l
eading to the quadratic assignment problem which is NP-hard. Several so called s
econd order methods have been proposed to solve this problem. In recent years hy
pergraph matching leading to a third order problem became popular as it allows f
or better integration of geometric information. For most of these third order al
gorithms no theoretical guarantees are known. In this paper we propose a general
 framework for tensor block coordinate ascent methods for hypergraph matching. W
e propose two algorithms which both come along with the guarantee of monotonic a
scent in the matching score on the set of discrete assignment matrices. In the e
xperiments we show that our new algorithms outperform previous work both in term
s of achieving better matching scores and matching accuracy. This holds in parti
cular for very challenging settings where one has a high number of outliers and
other forms of noise.
****

TILDE: A Temporally Invariant Learned DEtector
Yannick Verdie, Kwang Yi, Pascal Fua, Vincent Lepetit; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5279-5288
We introduce a learning-based approach to detect repeatable keypoints under dras
tic imaging changes of weather and lighting conditions to which state-of-the-art
 keypoint detectors are surprisingly sensitive. We first identify good keypoint
candidates in multiple training images taken from the same viewpoint. We then tr
ain a regressor to predict a score map whose maxima are those points so that the
y can be found by simple non-maximum suppression. As there are no standard data
sets to test the influence of these kinds of changes, we created our own, which
we will make publicly available. We will show that our method significantly outp
erforms the state-of-the-art methods in such challenging conditions, while still
 achieving state-of-the-art performance on the untrained standard Oxford dataset
.
****

A Maximum Entropy Feature Descriptor for Age Invariant Face Recognition
Dihong Gong, Zhifeng Li, Dacheng Tao, Jianzhuang Liu, Xuelong Li; Proceedings of
 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp
. 5289-5297
In this paper, we propose a new approach to overcome the representation and matc
hing problems in age invariant face recognition. First, a new maximum entropy fe
ature descriptor (MEFD) is developed that encodes the microstructure of facial i
mages into a set of discrete codes in terms of maximum entropy. By densely sampl
ing the encoded face image, sufficient discriminatory and expressive information
 can be extracted for further analysis. A new matching method is also developed,
 called identity factor analysis (IFA), to estimate the probability that two fac
es have the same underlying identity. The effectiveness of the framework is conf
irmed by extensive experimentation on two face aging datasets, MORPH (the larges
t public-domain face aging dataset) and FGNET. We also conduct experiments on th

e famous LFW dataset to demonstrate the excellent generalizability of our new ap
proach.
****

Sense Discovery via Co-Clustering on Images and Text
Xinlei Chen, Alan Ritter, Abhinav Gupta, Tom Mitchell; Proceedings of the IEEE C
onference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5298-5306
We present a co-clustering framework that can be used to discover multiple seman
tic and visual senses of a given Noun Phrase (NP). Unlike traditional clustering
 approaches which assume a one-to-one mapping between the clusters in the text-b
ased feature space and the visual space, we adopt a one-to-many mapping between
the two spaces. This is primarily because each semantic sense (concept) can corr
espond to different visual senses due to viewpoint and appearance variations. Ou
r structure-EM style optimization not only extracts the multiple senses in both
semantic and visual feature space, but also discovers the mapping between the se
nses. We introduce a challenging dataset (CMU Polysemy-30) for this problem cons
isting of 30 NPs ($\sim$5600 labeled instances out of $\sim$22K total instances)
. We have also conducted a large-scale experiment that performs sense disambigua
tion for $\sim$2000 NPs.
****

An Approximate Shading Model for Object Relighting
Zicheng Liao, Kevin Karsch, David Forsyth; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5307-5314
We propose an approximate shading model for image-based object modeling and inse
rtion. Our approach is a hybrid of 3D rendering and image-based composition. It
avoids the difficulties of physically accurate shape estimation from a single im
age, and allows for more flexible image composition than pure image-based method
s. The model decomposes the shading field into (a) a rough shape term that can b
e reshaded, (b) a parametric shading detail that encodes missing features from t
he first term, and (c) a geometric detail term that captures fine-scale material
 properties. With this object model, we build an object relighting system that
allows an artist to select an object from an image and insert it into a 3D scene
. Through simple interactions, the system can adjust illumination on the inserte
d object so that it appears more naturally in the scene. Our quantitative evalua
tion and extensive user study suggest our method is a promising alternative to e
xisting methods of object insertion.
****

Deep Domain Adaptation for Describing People Based on Fine-Grained Clothing Attr
ibutes
Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M. Brown, Jian Dong, Shuicheng Yan
; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
(CVPR), 2015, pp. 5315-5324
We address the problem of describing people based on fine-grained clothing attri
butes. This is an important problem for many practical applications, such as ide
ntifying target suspects or finding missing people based on detailed clothing de
scriptions in surveillance videos or consumer photos. We approach this problem b
y first mining clothing images with fine-grained attribute labels from online sh
opping stores. A large-scale dataset is built with about one million images and
fine-detailed attribute sub-categories, such as various shades of color (e.g., w
atermelon red, rosy red, purplish red), clothing types (e.g., down jacket, denim
 jacket), and patterns (e.g., thin horizontal stripes, houndstooth). As these im
ages are taken in ideal pose/lighting/background conditions, it is unreliable to
 directly use them as training data for attribute prediction in the domain of un
constrained images captured, for example, by mobile phones or surveillance camer
as. In order to bridge this gap, we propose a novel double-path deep domain adap
tation network to model the data from the two domains jointly. Several alignment
 cost layers placed in-between the two columns ensure the consistency of the two
 domain features and the feasibility to predict unseen attribute categories in o
ne of the domains. Finally, to achieve a working system with automatic human bod
y alignment, we trained an enhanced RCNN-based detector to localize human bodies
 in images. Our extensive experimental evaluation demonstrates the effectiveness

of the proposed approach for describing people based on fine-grained clothing attributes.
****

A Convolutional Neural Network Cascade for Face Detection
Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5325-5334
In real-world face detection, large visual variations, such as those due to pose, expression, and lighting, demand an advanced discriminative model to accurately differentiate faces from the backgrounds. Consequently, effective models for the problem tend to be computationally prohibitive. To address these two conflicting challenges, we propose a cascade architecture built on convolutional neural networks (CNNs) with very powerful discriminative capability, while maintaining high performance. The proposed CNN cascade operates at multiple resolutions, quickly rejects the background regions in the fast low resolution stages, and carefully evaluates a small number of challenging candidates in the last high resolution stage. To improve localization effectiveness, and reduce the number of candidates at later stages, we introduce a CNN-based calibration stage after each of the detection stages in the cascade. The output of each calibration stage is used to adjust the detection window position for input to the subsequent stage. The proposed method runs at 14 FPS on a single CPU core for VGA-resolution images and 100 FPS using a GPU, and achieves state-of-the-art detection performance on two public face detection benchmarks.
****

Visual Vibrometry: Estimating Material Properties From Small Motion in Video
Abe Davis, Katherine L. Bouman, Justin G. Chen, Michael Rubinstein, Fredo Durand, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5335-5343
The estimation of material properties is important for scene understanding, with many applications in vision, robotics, and structural engineering. This paper connects fundamentals of vibration mechanics with computer vision techniques in order to infer material properties from small, often imperceptible motion in video. Objects tend to vibrate in a set of preferred modes. The shapes and frequencies of these modes depend on the structure and material properties of an object. Focusing on the case where geometry is known or fixed, we show how information about an object's modes of vibration can be extracted from video and used to make inferences about that object's material properties. We demonstrate our approach by estimating material properties for a variety of rods and fabrics by passively observing their motion in high-speed and regular framerate video.
****

Jointly Learning Heterogeneous Features for RGB-D Activity Recognition
Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, Jianguo Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5344-5352
In this paper, we focus on heterogeneous feature learning for RGB-D activity recognition. Considering that features from different channels could share some similar hidden structures, we propose a joint learning model to simultaneously explore the shared and feature-specific components as an instance of heterogenous multi-task learning. The proposed model in an unified framework is capable of: 1) jointly mining a set of subspaces with the same dimensionality to enable the multi-task classifier learning, and 2) meanwhile, quantifying the shared and feature-specific components of features in the subspaces. To efficiently train the joint model, a three-step iterative optimization algorithm is proposed, followed by two inference models. Extensive results on three activity datasets have demonstrated the efficacy of the proposed method. In addition, a novel RGB-D activity dataset focusing on human-object interaction is collected for evaluating the proposed method, which will be made available to the community for RGB-D activity benchmarking and analysis.
****

Convolutional Neural Networks at Constrained Time Cost

Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5353-5360

Though recent advanced convolutional neural networks (CNNs) have been improving the image recognition accuracy, the models are getting more complex and time-consuming. For real-world applications in industrial and commercial scenarios, engineers and developers are often faced with the requirement of constrained time budget. In this paper, we investigate the accuracy of CNNs under constrained time cost. Under this constraint, the designs of the network architectures should exhibit as trade-offs among the factors like depth, numbers of filters, filter sizes, etc. With a series of controlled comparisons, we progressively modify a baseline model while preserving its time complexity. This is also helpful for understanding the importance of the factors in network designs. We present an architecture that achieves very competitive accuracy in the ImageNet dataset (11.8% top-5 error, 10-view test), yet is 20% faster than ``AlexNet'' (16.0% top-5 error, 10-view test).

****

Fine-Grained Histopathological Image Analysis via Robust Segmentation and Large-Scale Retrieval

Xiaofan Zhang, Hai Su, Lin Yang, Shaoting Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5361-5368

Computer-aided diagnosis of medical images requires thorough analysis of image details. For example, examining all cells enables fine-grained categorization of histopathological images. Traditional computational methods may have efficiency issues when performing such detailed analysis. In this paper, we propose a robust and scalable solution to achieve this. Specifically, a robust segmentation method is developed to delineate region-of-interests (e.g., cells) accurately, using hierarchical voting and repulsive active contour. A hashing-based large-scale retrieval approach is also designed to examine and classify them by comparing with a massive training database. We evaluate this proposed framework on a challenging and important clinical use case, i.e., differentiation of two types of lung cancers (the adenocarcinoma and the squamous carcinoma), using thousands of histopathological images extracted from hundreds of patients. Our method has achieved promising performance, i.e., 87.3% accuracy and 1.68 seconds by searching among half-million cells.

****

L0TV: A New Method for Image Restoration in the Presence of Impulse Noise

Ganzhao Yuan, Bernard Ghanem; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5369-5377

Total Variation (TV) is an effective and popular prior model in the field of regularization- image processing. This paper focuses on TV for image restoration in the presence of impulse noise. This type of noise frequently arises in data acquisition and transmission due to many reasons, e.g. a faulty sensor or analog-to-digital converter errors. Removing this noise is an important task in image restoration. State-of-the-art methods such as Adaptive Outlier Pursuit(AOP) [42], which is based on TV with L02-norm data fidelity, only give sub-optimal performance. In this paper, we propose a new method, called L0TV-PADMM, which solves the TV-based restoration problem with L0-norm data fidelity. To effectively deal with the resulting non-convex non-smooth optimization problem, we first reformulate it as an equivalent MPEC (Mathematical Program with Equilibrium Constraints), and then solve it using a proximal Alternating Direction Method of Multipliers (PADMM). Our L0TV-PADMM method finds a desirable solution to the original L0-norm optimization problem and is proven to be convergent under mild conditions. We apply L0TV-PADMM to the problems of image denoising and deblurring in the presence of impulse noise. Our extensive experiments demonstrate that L0TV-PADMM outperforms state-of-the-art image restoration methods.

****

Modeling Video Evolution for Action Recognition

Basura Fernando, Efstratios Gavves, Jose Oramas M., Amir Ghodrati, Tinne Tuytelaars; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5378-5387

In this paper we present a method to capture video-wide temporal information for action recognition. We postulate that a function capable of ordering the frames of a video temporally (based on the appearance) captures well the evolution of the appearance within the video. We learn such ranking functions per video via a ranking machine and use the parameters of these as a new video representation. The proposed method is easy to interpret and implement, fast to compute and effective in recognizing a wide variety of actions. We perform a large number of evaluations on datasets for generic action recognition (Hollywood2 and HMDB51), fine-grained actions (MPII- cooking activities) and gestures (Chalearn). Results show that the proposed method brings an absolute improvement of 7-10\%, while being compatible with and complementary to further improvements in appearance and local motion based methods.
****

Long-Term Correlation Tracking
Chao Ma, Xiaokang Yang, Chongyang Zhang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5388-5396
In this paper, we address the problem of long-term visual tracking where the target objects undergo significant appearance variation due to deformation, abrupt motion, heavy occlusion and out-of-the-view. In this setting, we decompose the task of tracking into translation and scale estimation of objects. We show that the correlation between temporal context considerably improves the accuracy and reliability for translation estimation, and it is effective to learn the discriminative correlation filters from the most confident frames to estimate the scale change. In addition, we train an online random fern classifier to re-detect objects in case of tracking failure. Extensive experimental results on large-scale benchmark datasets show that the proposed algorithm performs favorably against the state-of-the-art methods in terms of efficiency, accuracy, and robustness.
****

Joint Tracking and Segmentation of Multiple Targets
Anton Milan, Laura Leal-Taixe, Konrad Schindler, Ian Reid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5397-5406
Tracking-by-detection has proven to be the most successful strategy to address the task of tracking multiple targets in unconstrained scenarios. Traditionally, a set of sparse detections, generated in a preprocessing step, serves as input to a high-level tracker whose goal is to correctly associate these "dots" over time. An obvious shortcoming of this approach is that most information available in image sequences is simply ignored by thresholding weak detection responses and applying non-maximum suppression. We propose a multi-target tracker that exploits low level image information and associates every (super)-pixel to a specific target or classifies it as background. As a result, we obtain an video segmentation in addition to the classical bounding-box representation in unconstrained, real-world sequences. Our method shows encouraging results on many standard benchmark sequences and significantly outperforms state-of-the-art tracking-by-detection approaches in crowded scenes with long-term partial occlusions.
****

RGBD-Fusion: Real-Time High Precision Depth Recovery
Roy Or - El, Guy Rosman, Aaron Wetzler, Ron Kimmel, Alfred M. Bruckstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5407-5416
The popularity of low-cost RGB-D scanners is increasing on a daily basis. Nevertheless, existing scanners often cannot capture subtle details in the environment. We present a novel method to enhance the depth map by fusing the intensity and depth information to create more detailed range profiles. The lighting model we use can handle natural scene illumination. It is integrated in a shape from shading like technique to improve the visual fidelity of the reconstructed object. Unlike previous efforts in this domain, the detailed geometry is calculated directly, without the need to explicitly find and integrate surface normals. In addition, the proposed method operates four orders of magnitude faster than the stat

e of the art. Qualitative and quantitative visual and statistical evidence suppo
rt the improvement in the depth obtained by the suggested method.
****

Modeling Deformable Gradient Compositions for Single-Image Super-Resolution
Yu Zhu, Yanning Zhang, Boyan Bonev, Alan L. Yuille; Proceedings of the IEEE Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5417-5425
We propose a single-image super-resolution method based on the gradient reconstr
uction. To predict the gradient field, we collect a dictionary of gradient patte
rns from an external set of images. We observe that there are patches representi
ng singular primitive structures (e.g. a single edge), and non-singular ones (e.
g. a triplet of edges). Based on the fact that singular primitive patches are mo
re invariant to the scale change (i.e. have less ambiguity across different scal
es), we represent the non-singular primitives as compositions of singular ones,
each of which is allowed some deformation. Both the input patches and dictionary
 elements are decomposed to contain only singular primitives. The compositional
aspect of the model makes the gradient field more reliable. The deformable aspec
t makes the dictionary more expressive. As shown in our experimental results, th
e proposed method outperforms the state-of-the-art methods.
****

Generalized Video Deblurring for Dynamic Scenes
Tae Hyun Kim, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Visi
on and Pattern Recognition (CVPR), 2015, pp. 5426-5434
Several state-of-the-art video deblurring methods are based on a strong assumpti
on that the captured scenes are static. These methods fail to deblur blurry vide
os in dynamic scenes. We propose a video deblurring method to deal with general
blurs inherent in dynamic scenes, contrary to other methods. To handle locally v
arying and general blurs caused by various sources, such as camera shake, moving
 objects, and depth variation in a scene, we approximate pixel-wise kernel with
bidirectional optical flows. Therefore, we propose a single energy model that si
multaneously estimates optical flows and latent frames to solve our deblurring p
roblem. We also provide a framework and efficient solvers to optimize the energy
 model. By minimizing the proposed energy function, we achieve significant impro
vements in removing blurs and estimating accurate optical flows in blurry frames
. Extensive experimental results demonstrate the superiority of the proposed met
hod in real and challenging videos that state-of-the-art methods fail in either
deblurring or optical flow estimation.
****

Active Pictorial Structures
Epameinondas Antonakos, Joan Alabort-i-Medina, Stefanos Zafeiriou; Proceedings o
f the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, p
p. 5435-5444
In this paper we present a novel generative deformable model motivated by Pictor
ial Structures (PS) and Active Appearance Models (AAMs) for object alignment in-
the-wild. Inspired by the tree structure used in PS, the proposed Active Pictori
al Structures (APS) model the appearance of the object using multiple graph-base
d pairwise normal distributions (Gaussian Markov Random Field) between the patch
es extracted from the regions around adjacent landmarks. We show that this formu
lation is more accurate than using a single multivariate distribution (Principal
 Component Analysis) as commonly done in the literature. APS employ a weighted i
nverse compositional Gauss-Newton optimization with fixed Jacobian and Hessian t
hat achieves close to real-time performance and state-of-the-art results. Finall
y, APS have a spring-like graph-based deformation prior term that makes them rob
ust to bad initializations. We present extensive experiments on the task of face
 alignment, showing that APS outperform current state-of-the-art methods. To the
 best of our knowledge, the proposed method is the first weighted inverse compos
itional technique that proves to be so accurate and efficient at the same time.
****

Ego-Surfing First-Person Videos
Ryo Yonetani, Kris M. Kitani, Yoichi Sato; Proceedings of the IEEE Conference on
 Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5445-5454

We envision a future time when wearable cameras (e.g., small cameras in glasses or pinned on a shirt collar) are worn by the masses and record first-person point-of-view (POV) videos of everyday life. While these cameras can enable new assistive technologies and novel research challenges, they also raise serious privacy concerns. For example, first-person videos passively recorded by wearable cameras will necessarily include anyone who comes into the view of a camera -- with or without consent. Motivated by these benefits and risks, we develop a self-search technique tailored to first-person POV videos. The key observation of our work is that the egocentric head motions of a target person (i.e., the self) are observed both in the POV video of the target and observer. The motion correlation between the target person's video and the observer's video can then be used to uniquely identify instances of the self. We incorporate this feature into our proposed approach that computes the motion correlation over supervoxel hierarchies to localize target instances in observer videos. Our proposed approach significantly improves self-search performance over several well-known face detectors and recognizers. Furthermore, we show how our approach can enable several practical applications such as privacy filtering, automated video collection and social group discovery.
****

Visual Saliency Based on Multiscale Deep Features
Guanbin Li, Yizhou Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5455-5463
Visual saliency is a fundamental problem in both cognitive and computational sciences, including computer vision. In this paper, we discover that a high-quality visual saliency model can be learned from multiscale features extracted using deep convolutional neural networks(CNN), which have had many successes in visual recognition tasks. For learning such saliency models, we introduce a neural network architecture, which has fully connected layers on top of CNNs responsible for extracting features at three different scales. Our learned saliency model is capable of achieving state-of-the-art performance on all public benchmarks. We then propose a refinement method to enhance the spatial coherence of our saliency results. Finally, we point out that aggregating multiple saliency maps computed for different levels of image segmentation can further boost the performance, yielding saliency maps better than those generated from a single region decomposition. To promote further research and evaluation of visual saliency models, we also construct a large database of 4447 challenging images and their pixelwise saliency annotation. Experimental results demonstrate that our proposed method significantly outperforms all existing saliency estimation techniques, improving the F-Measure by 5.0% and 13.2% respectively on the MSRA-B dataset and our new dataset, and lowering the mean absolute error by 5.7% and 35.1% respectively on the same two datasets.
****

Recovering Inner Slices of Translucent Objects by Multi-Frequency Illumination
Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5464-5472
This paper describes a method for recovering appearance of inner slices of translucent objects. The outer appearance of translucent objects is a summation of the appearance of slices at all depths, where each slice is blurred by depth-dependent point spread functions (PSFs). By exploiting the difference of low-pass characteristics of depth-dependent PSFs, we develop a multi-frequency illumination method for obtaining the appearance of individual inner slices using a coaxial projector-camera setup. Specifically, by measuring the target object with varying the spatial frequency of checker patterns emitted from a projector, our method recovers inner slices via a simple linear solution method. We quantitatively evaluate accuracy of the proposed method by simulations and show qualitative recovery results using real-world scenes.
****

Local High-Order Regularization on Data Manifolds
Kwang In Kim, James Tompkin, Hanspeter Pfister, Christian Theobalt; Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5473-5481
The common graph Laplacian regularizer is well-established in semi-supervised learning and spectral dimensionality reduction. However, as a first-order regularizer, it can lead to degenerate functions in high-dimensional manifolds. The iterated graph Laplacian enables high-order regularization, but it has a high computational complexity and so cannot be applied to large problems. We introduce a new regularizer which is globally high order and so does not suffer from the degeneracy of the graph Laplacian regularizer, but is also sparse for efficient computation in semi-supervised learning applications. We reduce computational complexity by building a local first-order approximation of the manifold as a surrogate geometry, and construct our high-order regularizer based on local derivative evaluations therein. Experiments on human body shape and pose analysis demonstrate the effectiveness and efficiency of our method.
****
Fine-Grained Classification of Pedestrians in Video: Benchmark and State of the Art
David Hall, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5482-5491
A video dataset that is designed to study fine-grained categorisation of pedestrians is introduced. Pedestrians were recorded ``in-the-wild'' from a moving vehicle. Annotations include bounding boxes, tracks, 14 keypoints with occlusion information and the fine-grained categories of age (5 classes), sex (2 classes), weight (3 classes) and clothing style (4 classes). There are a total of 27,454 bounding box and pose labels across 4222 tracks. This dataset is designed to train and test algorithms for fine-grained categorisation of people; it is also useful for benchmarking tracking, detection and pose estimation of pedestrians. State-of-the-art algorithms for fine-grained classification and pose estimation were tested using the dataset and the results are reported as a useful performance baseline.
****
Curriculum Learning of Multiple Tasks
Anastasia Pentina, Viktoriia Sharmanska, Christoph H. Lampert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5492-5500
Sharing information between multiple tasks enables algorithms to achieve good generalization performance even from small amounts of training data. However, in a realistic scenario of multi-task learning not all tasks are equally related to each other, hence it could be advantageous to transfer information only between the most related tasks. In this work we propose an approach that processes multiple tasks in a sequence with sharing between subsequent tasks instead of solving all tasks jointly. Subsequently, we address the question of curriculum learning of tasks, i.e. finding the best order of tasks to be learned. Our approach is based on a generalization bound criterion for choosing the task order that optimizes the average expected classification performance over all tasks. Our experimental results show that learning multiple related tasks sequentially can be more effective than learning them jointly, the order in which tasks are being solved affects the overall performance, and that our model is able to automatically discover a favourable order of tasks.
****
How Many Bits Does it Take for a Stimulus to Be Salient?
Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V. Bajic, Yufeng Shan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5501-5510
Visual saliency has been shown to depend on the unpredictability of the visual stimulus given its surround. Various previous works have advocated the equivalence between stimulus saliency and uncompressibility. We propose a direct measure of this quantity, namely the number of bits required by an optimal video compressor to encode a given video patch, and show that features derived from this measure are highly predictive of eye fixations. To account for global saliency effect

s, these are embedded in a Markov random field model. The resulting saliency mea sure is shown to achieve state-of-the-art accuracy for the prediction of fixatio ns, at a very low computational cost. Since most modern cameras incorporate vide o encoders, this paves the way for in-camera saliency estimation, which could be useful in a variety of computer vision applications.
****

Discrete Optimization of Ray Potentials for Semantic 3D Reconstruction
Nikolay Savinov, Lubor Ladicky, Christian Hane, Marc Pollefeys; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5511-5518
Dense semantic 3D reconstruction is typically formulated as a discrete or contin uous problem over label assignments in a voxel grid, combining semantic and dept h likelihoods in a Markov Random Field framework. The depth and semantic informa tion is incorporated as a unary potential, smoothed by a pairwise regularizer. H owever, modelling likelihoods as a unary potential does not model the problem co rrectly leading to various undesirable visibility artifacts. We propose to form ulate an optimization problem that directly optimizes the reprojection error of the 3D model with respect to the image estimates, which corresponds to the optim ization over rays, where the cost function depends on the semantic class and dep th of the first occupied voxel along the ray. The 2-label formulation is made fe asible by transforming it into a graph-representable form under QPBO relaxation, solvable using graph cut. The multi-label problem is solved by applying $\alpha $-expansion using the same relaxation in each expansion move. Our method was ind eed shown to be feasible in practice, running comparably fast to the competing m ethods, while not suffering from ray potential approximation artifacts.
****

SOLD: Sub-Optimal Low-rank Decomposition for Efficient Video Segmentation
Chenglong Li, Liang Lin, Wangmeng Zuo, Shuicheng Yan, Jin Tang; Proceedings of t he IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5519-5527
This paper investigates how to perform robust and efficient unsupervised video s egmentation while suppressing the effects of data noises and/or corruptions. We propose a general algorithm, called Sub-Optimal Low-rank Decomposition (SOLD), w hich pursues the low-rank representation for video segmentation. Given the super voxels affinity matrix of an observed video sequence, SOLD seeks a sub-optimal s olution by making the matrix rank explicitly determined. In particular, the affi nity matrix with the rank fixed can be decomposed into two sub-matrices of low r ank, and then we iteratively optimize them with closed-form solutions. Moreover, we incorporate a discriminative replication prior into our framework based on t he obervation that small-size video patterns tend to recur frequently within the same object. The video can be segmented into several spatio-temporal regions by applying the Normalized-Cut (NCut) algorithm with the solved low-rank represent ation. To process the streaming videos, we apply our algorithm sequentially over a batch of frames over time, in which we also develop several temporal consiste nt constraints improving the robustness. Extensive experiments on the public ben chmarks demonstrate superior performance of our framework over other state-of-th e-art approaches.
****

On the Appearance of Translucent Edges
Ioannis Gkioulekas, Bruce Walter, Edward H. Adelson, Kavita Bala, Todd Zickler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (C VPR), 2015, pp. 5528-5536
Edges in images of translucent objects are very different from edges in images o f opaque objects. The physical causes for these differences are hard to characte rize analytically and are not well understood. This paper considers one class of translucency edges---those caused by a discontinuity in surface orientation---a nd describes the physical causes of their appearance. We simulate thousands of t ranslucency edge profiles using many different scattering material parameters, a nd we explain the resulting variety of edge patterns by qualitatively analyzing light transport. We also discuss the existence of shape and material metamers, o

r combinations of distinct shape or material parameters that generate the same e
dge profile. This knowledge is relevant to visual inference tasks that involve t
ranslucent objects, such as shape or material estimation.
****

On Pairwise Costs for Network Flow Multi-Object Tracking
Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, Josef Sivic; Proceedings of the
 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 55
37-5545
Multi-object tracking has been recently approached with the min-cost network flo
w optimization techniques. Such methods simultaneously resolve multiple object t
racks in a video and enable modeling of dependencies among tracks. Min-cost netw
ork flow methods also fit well within the "tracking-by-detection" paradigm where
 object trajectories are obtained by connecting per-frame outputs of an object d
etector. Object detectors, however, often fail due to occlusions and clutter in
the video. To cope with such situations, we propose to add pairwise costs to the
 min-cost network flow framework. While integer solutions to such a problem beco
me NP-hard, we design a convex relaxation solution with an efficient rounding he
uristic which empirically gives certificates of small suboptimality. We evaluate
 two particular types of pairwise costs and demonstrate improvements over recent
 tracking methods in real-world video sequences.
****

Fine-Grained Recognition Without Part Annotations
Jonathan Krause, Hailin Jin, Jianchao Yang, Li Fei-Fei; Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5546-555
5
Scaling up fine-grained recognition to all domains of fine-grained objects is a
challenge the computer vision community will need to face in order to realize it
s goal of recognizing all object categories. Current state-of-the-art techniques
 rely heavily upon the use of keypoint or part annotations, but scaling up to hu
ndreds or thousands of domains renders this annotation cost-prohibitive for all
but the most important categories. In this work we propose a method for fine-gra
ined recognition that uses no part annotations. Our method is based on generatin
g parts using co-segmentation and alignment, which we combine in a discriminativ
e mixture. Experimental results show its efficacy, demonstrating state-of-the-ar
t results even when compared to methods that use part annotations during trainin
g.
****

Robust Reconstruction of Indoor Scenes
Sungjoon Choi, Qian-Yi Zhou, Vladlen Koltun; Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5556-5565
We present an approach to indoor scene reconstruction from RGB-D video. The key
idea is to combine geometric registration of scene fragments with robust global
optimization based on line processes. Geometric registration is error-prone due
to sensor noise, which leads to aliasing of geometric detail and inability to di
sambiguate different surfaces in the scene. The presented optimization approach
disables erroneous geometric alignments even when they significantly outnumber c
orrect ones. Experimental results demonstrate that the presented approach substa
ntially increases the accuracy of reconstructed scene models.
****