Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images

Mateusz Malinowski, Marcus Rohrbach, Mario Fritz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1-9

We address a question answering task on real-world images that is set up as a Visual Turing Test. By combining latest advances in image representation and natural language processing, we propose Neural-Image-QA, an end-to-end formulation to this problem for which all parts are trained jointly. In contrast to previous efforts, we are facing a multi-modal problem where the language output (answer) is conditioned on visual and natural language input (image and question). Our approach Neural-Image-QA doubles the performance of the previous best approach on this problem. We provide additional insights into the problem by analyzing how much information is contained only in the language part for which we provide a new human baseline. To study human consensus, which is related to the ambiguities inherent in this challenging task, we propose two novel metrics and collect additional answers which extends the original DAQUAR dataset to DAQUAR-Consensus.

****

Segment-Phrase Table for Semantic Segmentation, Visual Entailment and Paraphrasing

Hamid Izadinia, Fereshteh Sadeghi, Santosh K. Divvala, Hannaneh Hajishirzi, Yejin Choi, Ali Farhadi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 10-18

We introduce Segment-Phrase Table (SPT), a large collection of bijective associations between textual phrases and their corresponding segmentations. Leveraging recent progress in object recognition and natural language semantics, we show how we can successfully build a high-quality segment-phrase table using minimal human supervision. More importantly, we demonstrate the unique value unleashed by this rich bimodal resource, for both vision as well as natural language understanding. First, we show that fine-grained textual labels facilitate contextual reasoning that helps in satisfying semantic constraints across image segments. This feature enables us to achieve state-of-the-art segmentation results on benchmark datasets. Next, we show that the association of high-quality segmentations to textual phrases aids in richer semantic understanding and reasoning of these textual phrases. Leveraging this feature, we motivate the problem of visual entailment and visual paraphrasing, and demonstrate its utility on a large dataset.

****

Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 19-27

Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story. This paper aims to align books to their movie releases in order to provide rich descriptive explanations for visual content that go semantically far beyond the captions available in the current datasets. To align movies and books we propose a neural sentence embedding that is trained in an unsupervised way from a large corpus of books, as well as a video-text neural embedding for computing similarities between movie clips and sentences in the book. We propose a context-aware CNN to combine information from multiple sources. We demonstrate good quantitative performance for movie/book alignment and show several qualitative examples that showcase the diversity of tasks our model can be used for.

****

Learning Query and Image Similarities With Ranking Canonical Correlation Analysis

Ting Yao, Tao Mei, Chong-Wah Ngo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 28-36

One of the fundamental problems in image search is to learn the ranking functions, i.e., similarity between the query and image. The research on this topic has evolved through two paradigms: feature-based vector model and image ranker learn ing. The former relies on the image surrounding texts, while the latter learns a ranker based on human labeled query-image pairs. Each of the paradigms has its own limitation. The vector model is sensitive to the quality of text descriptions, and the learning paradigm is difficult to be scaled up as human labeling is always too expensive to obtain. We demonstrate in this paper that the above two limitations can be well mitigated by jointly exploring subspace learning and the use of click-through data. Specifically, we propose a novel Ranking Canonical Correlation Analysis (RCCA) for learning query and image similarities. RCCA initially finds a common subspace between query and image views by maximizing their correlations, and further simultaneously learns a bilinear query-image similarity function and adjusts the subspace to preserve the preference relations implicit in the click-through data. Once the subspace is finalized, query-image similarity can be computed by the bilinear similarity function on their mappings in this subspace. On a large-scale click-based image dataset with 11.7 million queries and one million images, RCCA is shown to be powerful for image search with superior performance over several state-of-the-art methods on both keyword-based and query-by-example tasks.

****

Learning to See by Moving

Pulkit Agrawal, Joao Carreira, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 37-45

The current dominant paradigm for feature learning in computer vision relies on training neural networks for the task of object recognition using millions of hand labelled images. Is it also possible to learn features for a diverse set of visual tasks using any other form of supervision? In biology, living organisms developed the ability of visual perception for the purpose of moving and acting in the world. Drawing inspiration from this observation, in this work we investigated if the awareness of egomotion(i.e. self motion) can be used as a supervisory signal for feature learning. As opposed to the knowledge of class labels, information about egomotion is freely available to mobile agents. We found that using the same number of training images, features learnt using egomotion as supervision compare favourably to features learnt using class-label as supervision on the tasks of scene recognition, object recognition, visual odometry and keypoint matching.

****

Object Detection Using Generalization and Efficiency Balanced Co-Occurrence Features

Haoyu Ren, Ze-Nian Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 46-54

In this paper, we propose a high-accuracy object detector based on co-occurrence features. Firstly, we introduce three kinds of local co-occurrence features constructed by the traditional Haar, LBP, and HOG respectively. Then the boosted detectors are learned, where each weak classifier corresponds to a local image region with a co-occurrence feature. In addition, we propose a Generalization and Efficiency Balanced (GEB) framework for boosting training. In the feature selection procedure, the discrimination ability, the generalization power, and the computation cost of the candidate features are all evaluated for decision. As a result, the boosted detector achieves both high accuracy and good efficiency. It also shows performance competitive with the state-of-the-art methods for pedestrian detection and general object detection tasks.

****

Mining And-Or Graphs for Graph Matching and Object Discovery

Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 55-63

This paper reformulates the theory of graph mining on the technical basis of graph matching, and extends its scope of applications to computer vision. Given a set of attributed relational graphs (ARGs), we propose to use a hierarchical And-Or Graph (AoG) to model the pattern of maximal-size common subgraphs embedded in the ARGs, and we develop a general method to mine the AoG model from the unlabeled ARGs. This method provides a general solution to the problem of mining hiera

rchical models from unannotated visual data without exhaustive search of objects. We apply our method to RGB/RGB-D images and videos to demonstrate its generality and the wide range of applicability. The code will be available at https://sites.google.com/site/quanshizhang/mining-and-or-graphs.
****

Pose Induction for Novel Object Categories

Shubham Tulsiani, Joao Carreira, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 64-72

We address the task of predicting pose for objects of unannotated object categories from a small seed set of annotated object classes. We present a generalized classifier that can reliably induce pose given a single instance of a novel category. In case of availability of a large collection of novel instances, our approach then jointly reasons over all instances to improve the initial estimates. We empirically validate the various components of our algorithm and quantitatively show that our method produces reliable pose estimates. We also show qualitative results on a diverse set of classes and further demonstrate the applicability of our system for learning shape models of novel object classes.
****

Dynamic Texture Recognition via Orthogonal Tensor Dictionary Learning

Yuhui Quan, Yan Huang, Hui Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 73-81

Dynamic textures (DTs) are video sequences with stationary properties, which exhibit repetitive patterns over space and time. This paper aims at investigating the sparse coding based approach to characterizing local DT patterns for recognition. Owing to the high dimensionality of DT sequences, existing dictionary learning algorithms are not suitable for our purpose due to their high computational costs as well as poor scalability. To overcome these obstacles, we proposed a structured tensor dictionary learning method for sparse coding, which learns a dictionary structured with orthogonality and separability. The proposed method is very fast and more scalable to high-dimensional data than the existing ones. In addition, based on the proposed dictionary learning method, a DT descriptor is developed, which has better adaptivity, discriminability and scalability than the existing approaches. These advantages are demonstrated by the experiments on multiple datasets.
****

Convolutional Channel Features

Bin Yang, Junjie Yan, Zhen Lei, Stan Z. Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 82-90

Deep learning methods are powerful tools but often suffer from expensive computation and limited flexibility. An alternative is to combine light-weight models with deep representations. As successful cases exist in several visual problems, a unified framework is absent. In this paper, we revisit two widely used approaches in computer vision, namely filtered channel features and Convolutional Neural Networks (CNN), and absorb merits from both by proposing an integrated method called Convolutional Channel Features (CCF). CCF transfers low-level features from pre-trained CNN models to feed the boosting forest model. With the combination of CNN features and boosting forest, CCF benefits from the richer capacity in feature representation compared with channel features, as well as lower cost in computation and storage compared with end-to-end CNN methods. We show that CCF serves as a good way of tailoring pre-trained CNN models to diverse tasks without fine-tuning the whole network to each task by achieving state-of-the-art performances in pedestrian detection, face detection, edge detection and object proposal generation.
****

Local Convolutional Features With Unsupervised Training for Image Retrieval

Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 91-99

Patch-level descriptors underlie several important computer vision tasks, such as stereo-matching or content-based image retrieval. We introduce a deep convol utional architecture that yields patch-level descriptors, as an alternative to the popular SIFT descriptor for image retrieval.   The proposed family of descriptors, called Patch-CKN, adapt the recently introduced Convolutional Kernel Network (CKN), an unsupervised framework to learn convolutional architectures. We present a comparison framework to benchmark current deep convolutional approaches along with Patch-CKN for both patch and image retrieval, including our novel ``RomePatches'' dataset. Patch-CKN descriptors yield competitive results compared to supervised CNN alternatives on patch and image retrieval.

****

RIDE: Reversal Invariant Descriptor Enhancement

Lingxi Xie, Jingdong Wang, Weiyao Lin, Bo Zhang, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 100-108

In many fine-grained object recognition datasets, image orientation (left/right) might vary from sample to sample. Since handcrafted descriptors such as SIFT are not reversal invariant, the stability of image representation based on them is consequently limited. A popular solution is to augment the datasets by adding a left-right reversed copy for each original image. This strategy improves recognition accuracy to some extent, but also brings the price of almost doubled time and memory consumptions. In this paper, we present RIDE (Reversal Invariant Descriptor Enhancement) for fine-grained object recognition. RIDE is a generalized algorithm which cancels out the impact of image reversal by estimating the orientation of local descriptors, and guarantees to produce the identical representation for an image and its left-right reversed copy. Experimental results reveal the consistent accuracy gain of RIDE with various types of descriptors. We also provide insightful discussions on the working mechanism of RIDE and its generalization to other applications.

****

Discrete Tabu Search for Graph Matching

Kamil Adamczewski, Yumin Suh, Kyoung Mu Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 109-117

Graph matching is a fundamental problem in computer vision. In this paper, we propose a novel graph matching algorithm based on tabu search. The proposed method solves graph matching problem by casting it into an equivalent weighted maximum clique problem of the corresponding association graph, which we further penalize through introducing negative weights. Subsequent tabu search optimization allows for overcoming the convention of using positive weights. The method's distinct feature is that it utilizes the history of search to make more strategic decisions while looking for the optimal solution, thus effectively escaping local optima and in practice achieving superior results. The proposed method, unlike the existing algorithms, enables direct optimization in the original discrete space while encouraging rather than artificially enforcing hard one-to-one constraint, thus resulting in better solution. The experiments demonstrate the robustness of the algorithm in a variety of settings, presenting the state-of-the-art results. The code is available at http://cv.snu.ac.kr/research/ DTSGM/

****

Discriminative Learning of Deep Convolutional Feature Point Descriptors

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, Francesc Moreno-Noguer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 118-126

Deep learning has revolutionalized image-level tasks such as classification, but patch-level tasks, such as correspondence, still rely on hand-crafted features, e.g. SIFT. In this paper we use Convolutional Neural Networks (CNNs) to learn discriminant patch representations and in particular train a Siamese network with pairs of (non-)corresponding patches. We deal with the large number of potential pairs with the combination of a stochastic sampling of the training set and an aggressive mining strategy biased towards patches that are hard to classify. By using the L2 distance during both training and testing we develop 128-D descriptors whose euclidean distances reflect patch similarity, and which can be used as a drop-in replacement for any task involving SIFT. We demonstrate consistent performance gains over the state of the art, and generalize well against scaling and rotation, perspective transformation, non-rigid deformation, and illumination changes. Our descriptors are efficient to compute and amenable to modern GPUs, and are publicly available.

****

Amodal Completion and Size Constancy in Natural Scenes

Abhishek Kar, Shubham Tulsiani, Joao Carreira, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 127-135

We consider the problem of enriching current object detection systems with veridical object sizes and relative depth estimates from a single image. There are several technical challenges to this, such as occlusions, lack of calibration data and the scale ambiguity between object size and distance. These have not been addressed in full generality in previous work. Here we propose to tackle these issues by building upon advances in object recognition and using recently created large-scale datasets. We first introduce the task of amodal bounding box completion, which aims to infer the the full extent of the object instances in the image. We then propose a probabilistic framework for learning category-specific object size distributions from available annotations and leverage these in conjunction with amodal completions to infer veridical sizes of objects in novel images. Finally, we introduce a focal length prediction approach that exploits scene recognition to overcome inherent scale ambiguities and demonstrate qualitative results on challenging real-world scenes.

****

Learning Where to Position Parts in 3D

Marco Pedersoli, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 136-144

A common issue in deformable object detection is finding a good way to position the parts. This issue is even more outspoken when considering detection and pose estimation for 3D objects, where parts should be placed in a three-dimensional space. Some methods extract the 3D shape of the object from 3D CAD models. This limits their applicability to categories for which such models are available. Others represent the object with a predefined and simple shape (e.g. a cuboid). This extends the applicability of the model, but in many cases the pre-defined shape is too simple to properly represent the object in 3D. In this paper we propose a new method for the detection and pose estimation of 3D objects, that does not use any 3D CAD model or other 3D information. Starting from a simple and general 3D shape, we learn in a weakly supervised manner the 3D part locations that best fit the training data. As this method builds on a iterative estimation of the part locations, we introduce several speedups to make the method fast enough for practical experiments. We evaluate our model for the detection and pose estimation of faces and cars. Our method obtains results comparable with the state of the art, it is faster than most of the other approaches and does not need any additional 3D information.

****

Query Adaptive Similarity Measure for RGB-D Object Recognition

Yanhua Cheng, Rui Cai, Chi Zhang, Zhiwei Li, Xin Zhao, Kaiqi Huang, Yong Rui; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 145-153

This paper studies the problem of improving the top-1 accuracy of RGB-D object recognition. Despite of the impressive top-5 accuracies achieved by existing methods, their top-1 accuracies are not very satisfactory. The reasons are in two-fold: (1) existing similarity measures are sensitive to object pose and scale changes, as well as intra-class variations; and (2) effectively fusing RGB and depth cues is still an open problem. To address these problems, this paper first proposes a new similarity measure based on dense matching, through which objects in comparison are warped and aligned, to better tolerate variations. Towards RGB and depth fusion, we argue that a constant and golden weight doesn't exist. The two modalities have varying contributions when comparing objects from different categories. To capture such a dynamic characteristic, a group of matchers equipped with various fusion weights is constructed, to explore the responses of dense matching under different fusion configurations. All the response scores are final ly merged following a learning-to-combination way, which provides quite good generalization ability in practice. The proposed approach win the best results on several public benchmarks, e.g., achieves 92.7% top-1 test accuracy on the Washington RGB-D object dataset, with a 5.1% improvement over the state-of-the-art.

****

Listening With Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines

Chao Sui, Mohammed Bennamoun, Roberto Togneri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 154-162

This paper presents a novel feature learning method for visual speech recognition using Deep Boltzmann Machines (DBM). Unlike all existing visual feature extraction techniques which solely extracts features from video sequences, our method is able to explore both acoustic information and visual information to learn a better visual feature representation in the training stage. During the test stage, instead of using both audio and visual signals, only the videos are used for generating the missing audio feature, and both the given visual and given audio features are used to obtain a joint representation. We carried out our experiments on a large scale audio-visual data corpus, and experimental results show that our proposed techniques outperforms the performance of the hadncrafted features and features learned by other commonly used deep learning techniques.

****

Cluster-Based Point Set Saliency

Flora Ponjou Tasse, Jiri Kosinka, Neil Dodgson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 163-171

We propose a cluster-based approach to point set saliency detection, a challenge since point sets lack topological information. A point set is first decomposed into small clusters, using fuzzy clustering. We evaluate cluster uniqueness and spatial distribution of each cluster and combine these values into a cluster saliency function. Finally, the probabilities of points belonging to each cluster are used to assign a saliency to each point. Our approach detects fine-scale salient features and uninteresting regions consistently have lower saliency values. We evaluate the proposed saliency model by testing our saliency-based keypoint detection against a 3D interest point detection benchmark. The evaluation shows that our method achieves a good balance between false positive and false negative error rates, without using any topological information.

****

A Comprehensive Multi-Illuminant Dataset for Benchmarking of the Intrinsic Image Algorithms

Shida Beigpour, Andreas Kolb, Sven Kunz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 172-180

In this paper, we provide a new, real photo dataset with precise ground-truth for intrinsic image research. Prior ground-truth datasets have been restricted to rather simple illumination conditions and scene geometries, or have been enhanced using image synthesis methods. The dataset provided in this paper is based on complex multi-illuminant scenarios under multi-colored illumination conditions and challenging cast shadows. We provide full per-pixel intrinsic ground-truth data for these scenarios, i.e. reflectance, specularity, shading, and illumination for scenes as well as preliminary depth information. Furthermore, we evaluate 3 state-of-the-art intrinsic image recovery methods, using our dataset.

****

PatchMatch-Based Automatic Lattice Detection for Near-Regular Textures

Siying Liu, Tian-Tsong Ng, Kalyan Sunkavalli, Minh N. Do, Eli Shechtman, Nathan Carr; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 181-189

In this work, we investigate the problem of automatically inferring the lattice structure of near-regular textures (NRT) in real-world images. Our technique leverages the PatchMatch algorithm for finding k-nearest-neighbor (kNN) correspondences in an image. We use these kNNs to recover an initial estimate of the 2D wallpaper basis vectors, and seed vertices of the texture lattice. We iteratively expand this lattice by solving an MRF optimization problem. We show that we can d iscretize the space of good solutions for the MRF using the kNNs, allowing us to efficiently and accurately optimize the MRF energy function using the Particle Belief Propagation algorithm. We demonstrate our technique on a benchmark NRT dataset containing a wide range of images with geometric and photometric variations, and show that our method clearly outperforms the state of the art in terms of both texel detection rate and texel localization score.

****

A Data-Driven Metric for Comprehensive Evaluation of Saliency Models

Jia Li, Changqun Xia, Yafei Song, Shu Fang, Xiaowu Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 190-198

In the past decades, hundreds of saliency models have been proposed for fixation prediction, along with dozens of evaluation metrics. However, existing metrics, which are often heuristically designed, may draw conflict conclusions in comparing saliency models. As a consequence, it becomes somehow confusing on the selection of metrics in comparing new models with state-of-the-arts. To address this problem, we propose a data-driven metric for comprehensive evaluation of saliency models. Instead of heuristically designing such a metric, we first conduct extensive subjective tests to find how saliency maps are assessed by the human-being. Based on the user data collected in the tests, nine representative evaluation metrics are directly compared by quantizing their performances in assessing saliency maps. Moreover, we propose to learn a data-driven metric by using Convolutional Neural Network. Compared with existing metrics, experimental results show that the data-driven metric performs the most consistently with the human-being in evaluating saliency maps as well as saliency models.

****

A Matrix Decomposition Perspective to Multiple Graph Matching

Junchi Yan, Hongteng Xu, Hongyuan Zha, Xiaokang Yang, Huanxi Liu, Stephen Chu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 199-207

Graph matching has a wide spectrum of real-world applications and in general is known NP-hard. In many vision tasks, one realistic problem arises for finding the global node mappings across a batch of corrupted weighted graphs. This paper is an attempt to connect graph matching, especially multi-graph matching to the matrix decomposition model and its relevant on-the-shelf convex optimization algorithms. Our method aims to extract the common inliers and their synchronized permutations from disordered weighted graphs in the presence of deformation and outliers. Under the proposed framework, several variants can be derived in the hope of accommodating to other types of noises. Experimental results on both synthetic data and real images empirically show that the proposed paradigm exhibits several interesting behaviors and in many cases performs competitively with the state-of-the-arts.

****

Fast and Effective L0 Gradient Minimization by Region Fusion

Rang M. H. Nguyen, Michael S. Brown; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 208-216

L_0 gradient minimization can be applied to an input signal to control the number of non-zero gradients. This is useful in reducing small gradients generally associated with signal noise, while preserving important signal features. In computer vision, L_0 gradient minimization has found applications in image denoising, 3D mesh denoising, and image enhancement. Minimizing the L_0 norm, however, is an NP-hard problem because of its non-convex property. As a result, existing methods rely on approximation strategies to perform the minimization. In this paper, we present a new method to perform L_0 gradient minimization that is fast and effective. Our method uses a descent approach based on region fusion that converges faster than other methods while providing a better approximation of the optimal L_0 norm. In addition, our method can be applied to both 2D images and 3D mesh topologies. The effectiveness of our approach is demonstrated on a number of examples.

****

Generic Promotion of Diffusion-Based Salient Object Detection

Peng Jiang, Nuno Vasconcelos, Jingliang Peng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 217-225

In this work, we propose a generic scheme to promote any diffusion-based salient object detection algorithm by original ways to re-synthesize the diffusion matrix and construct the seed vector. We first make a novel analysis of the working mechanism of the diffusion matrix, which reveals the close relationship between saliency diffusion and spectral clustering. Following this analysis, we propose to re-synthesize the diffusion matrix from the most discriminative eigenvectors after adaptive re-weighting. Further, we propose to generate the seed vector based on the readily available diffusion maps, avoiding extra computation for color-based seed search. As a particular instance, we use inverse normalized Laplacian matrix as the original diffusion matrix and promote the corresponding salient object detection algorithm, which leads to superior performance as experimentally demonstrated.

****

Nighttime Haze Removal With Glow and Multiple Light Colors

Yu Li, Robby T. Tan, Michael S. Brown; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 226-234

This paper focuses on dehazing nighttime images. Most existing dehazing methods use models that are formulated to describe haze in daytime. Daytime models assume a single uniform light color attributed to a light source not directly visible in the scene. Nighttime scenes, however, commonly include visible lights sources with varying colors. These light sources also often introduce noticeable amounts of glow that is not present in daytime haze. To address these effects, we introduce a new nighttime haze model that accounts for the varying light sources and their glow. Our model is a linear combination of three terms: the direct transmission, airlight and glow. The glow term represents light from the light sources that is scattered around before reaching the camera. Based on the model, we propose a framework that first reduces the effect of the glow in the image, resulting in a nighttime image that consists of direct transmission and airlight only. We then compute a spatially varying atmospheric light map that encodes light colors locally. This atmospheric map is used to predict the transmission, which we use to obtain our nighttime scene reflection image. We demonstrate the effectiveness of our nighttime haze model and correction method on a number of examples and compare our results with existing daytime and nighttime dehazing methods' results.

****

Conformal and Low-Rank Sparse Representation for Image Restoration

Jianwei Li, Xiaowu Chen, Dongqing Zou, Bo Gao, Wei Teng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 235-243

Obtaining an appropriate dictionary is the key point when sparse representation is applied to computer vision or image processing problems such as image restoration. It is expected that preserving data structure during sparse coding and dictionary learning can enhance the recovery performance. However, many existing dictionary learning methods handle training samples individually, while missing relationships between samples, which result in dictionaries with redundant atoms but poor representation ability. In this paper, we propose a novel sparse representation approach called conformal and low-rank sparse representation (CLRSR) for image restoration problems. To achieve a more compact and representative dictionary, conformal property is introduced by preserving the angles of local geometry formed by neighboring samples in the feature space. Furthermore, imposing low-rank constraint on the coefficient matrix can lead more faithful subspaces and capture the global structure of data. We apply our CLRSR model to several image restoration tasks to demonstrate the effectiveness.

****

Patch Group Based Nonlocal Self-Similarity Prior Learning for Image Denoising

Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, Xiangchu Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 244-252

Patch based image modeling has achieved a great success in low level vision such as image denoising. In particular, the use of image nonlocal self-similarity (N

SS) prior, which refers to the fact that a local patch often has many nonlocal similar patches to it across the image, has significantly enhanced the denoising performance. However, in most existing methods only the NSS of input degraded image is exploited, while how to utilize the NSS of clean natural images is still an open problem. In this paper, we propose a patch group (PG) based NSS prior learning scheme to learn explicit NSS models from natural images for high performance denoising. PGs are extracted from training images by putting nonlocal similar patches into groups, and a PG based Gaussian Mixture Model (PG-GMM) learning algorithm is developed to learn the NSS prior. We demonstrate that, owe to the learned PG-GMM, a simple weighted sparse coding model, which has a closed-form solution, can be used to perform image denoising effectively, resulting in high PSNR measure, fast speed, and particularly the best visual quality among all competing methods.

****

Automatic Thumbnail Generation Based on Visual Representativeness and Foreground Recognizability

Jingwei Huang, Huarong Chen, Bin Wang, Stephen Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 253-261

We present an automatic thumbnail generation technique based on two essential considerations: how well they visually represent the original photograph, and how well the foreground can be recognized after the cropping and downsizing steps of thumbnailing. These factors, while important for the image indexing purpose of thumbnails, have largely been ignored in previous methods, which instead are designed to highlight salient content while disregarding the effects of downsizing. We propose a set of image features for modeling these two considerations of thumbnails, and learn how to balance their relative effects on thumbnail generation through training on image pairs composed of photographs and their corresponding thumbnails created by an expert photographer. Experiments show the effectiveness of this approach on a variety of images, as well as its advantages over related techniques.

****

SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks

Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 262-270

Saliency in Context (SALICON) is an ongoing effort that aims at understanding and predicting visual attention. Conventional saliency models typically rely on low-level image statistics to predict human fixations. While these models perform significantly better than chance, there is still a large gap between model prediction and human behavior. This gap is largely due to the limited capability of models in predicting eye fixations with strong semantic content, the so-called semantic gap. This paper presents a focused study to narrow the semantic gap with an architecture based on Deep Neural Network (DNN). It leverages the representational power of high-level semantics encoded in DNNs pretrained for object recognition. Two key components are fine-tuning the DNNs fully convolutionally with an objective function based on the saliency evaluation metrics, and integrating information at different image scales. We compare our method with 14 saliency models on 6 public eye tracking benchmark datasets. Results demonstrate that our DNNs can automatically learn features particularly for saliency prediction that surpass by a big margin the state-of-the-art. In addition, our model ranks top to date under all seven metrics on the MIT300 challenge set.

****

A Novel Sparsity Measure for Tensor Recovery

Qian Zhao, Deyu Meng, Xu Kong, Qi Xie, Wenfei Cao, Yao Wang, Zongben Xu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 271-279

In this paper, we propose a new sparsity regularizer for measuring the low-rank structure underneath a tensor. The proposed sparsity measure has a natural physical meaning which is intrinsically the size of the fundamental Kronecker basis to express the tensor. By embedding the sparsity measure into the tensor completi

on and tensor robust PCA frameworks, we formulate new models to enhance their capability in tensor recovery. Through introducing relaxation forms of the proposed sparsity measure, we also adopt the alternating direction method of multipliers (ADMM) for solving the proposed models. Experiments implemented on synthetic and multispectral image data sets substantiate the effectiveness of the proposed methods.
****

Oriented Object Proposals
Shengfeng He, Rynson W.H. Lau; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 280-288

In this paper, we propose a new approach to generate oriented object proposals (OOPs) to reduce the detection error caused by various orientations of the object. To this end, we propose to efficiently locate object regions according to pixelwise object probability, rather than measuring the objectness from a set of sampled windows. We formulate the proposal generation problem as a generative probabilistic model such that object proposals of different shapes (i.e., sizes and orientations) can be produced by locating the local maximum likelihoods. The new approach has three main advantages. First, it helps the object detector handle objects of different orientations. Second, as the shapes of the proposals may vary to fit the objects, the resulting proposals are tighter than the sampling windows with fixed sizes. Third, it avoids massive window sampling, and thereby reducing the number of proposals while maintaining a high recall. Experiments on the PASCAL VOC 2007 dataset show that the proposed OOP outperforms the state-of-the-art fast methods. Further experiments show that the rotation invariant property helps a class-specific object detector achieve better performance than the state-of-the-art proposal generation methods in either object rotation scenarios or general scenarios. Generating OOPs is very fast and takes only 0.5s per image.
****

Learning Nonlinear Spectral Filters for Color Image Reconstruction
Michael Moeller, Julia Diebold, Guy Gilboa, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 289-297

This paper presents the idea of learning optimal filters for color image reconstruction based on a novel concept of nonlinear spectral image decompositions recently proposed by Guy Gilboa. We use a multiscale image decomposition approach based on total variation regularization and Bregman iterations to represent the input data as the sum of image layers containing features at different scales. Filtered images can be obtained by weighted linear combinations of the different frequency layers. We introduce the idea of learning optimal filters for the task of image denoising, and propose the idea of mixing high frequency components of different color channels. Our numerical experiments demonstrate that learning the optimal weights can significantly improve the results in comparison to the standard variational approach, and achieves state-of-the-art image denoising results.
****

Beyond White: Ground Truth Colors for Color Constancy Correction
Dongliang Cheng, Brian Price, Scott Cohen, Michael S. Brown; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 298-306

A limitation in color constancy research is the inability to establish ground truth colors for evaluating corrected images. Many existing datasets contain images of scenes with a color chart included; however, only the chart's neutral colors (grayscale patches) are used to provide the ground truth for illumination estimation and correction. This is because the corrected neutral colors are known to lie along the achromatic line in the camera's color space (i.e. R=G=B) ; the correct RGB values of the other color patches are not known. As a result, most methods estimate a 3*3 diagonal matrix that ensures only the neutral colors are correct. In this paper, we describe how to overcome this limitation.   Specifically, we show that under certain illuminations, a diagonal 3*3 matrix is capable of correcting not only neutral colors, but all the colors in a scene.   This finding allows us to find the ground truth RGB values for the color chart in the camera's color space. We show how to use this information to correct all the image s in existing datasets to have correct colors. Working from these new color corrected datasets, we describe how to modify existing color constancy algorithms to perform better image correction.

****

RGB-Guided Hyperspectral Image Upsampling

Hyeokhyen Kwon, Yu-Wing Tai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 307-315

Hyperspectral imaging usually lack of spatial resolution due to limitations of hardware design of imaging sensors. On the contrary, latest imaging sensors capture a RGB image with resolution of multiple times larger than a hyperspectral image. In this paper, we present an algorithm to enhance and upsample the resolution of hyperspectral images. Our algorithm consists of two stages: spatial upsampling stage and spectrum substitution stage. The spatial upsampling stage is guided by a high resolution RGB image of the same scene, and the spectrum substitution stage utilizes sparse coding to locally refine the upsampled hyperspectral image through dictionary substitution. Experiments show that our algorithm is highly effective and has outperformed state-of-the-art matrix factorization based approaches.

****

Projection Onto the Manifold of Elongated Structures for Accurate Extraction

Amos Sironi, Vincent Lepetit, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 316-324

Detection of elongated structures in 2D images and 3D image stacks is a critical prerequisite in many applications and Machine Learning-based approaches have recently been shown to deliver superior performance. However, these methods essentially classify individual locations and do not explicitly model the strong relationship that exists between neighboring ones. As a result, isolated erroneous responses, discontinuities, and topological errors are present in the resulting score maps. We solve this problem by projecting patches of the score map to their nearest neighbors in a set of ground truth training patches. Our algorithm induces global spatial consistency on the classifier score map and returns results that are provably geometrically consistent. We apply our algorithm to challenging datasets in four different domains and show that it compares favorably to state-of-the-art methods.

****

Naive Bayes Super-Resolution Forest

Jordi Salvador, Eduardo Perez-Pellitero; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 325-333

This paper presents a fast, high-performance method for super resolution with external learning. The first contribution leading to the excellent performance is a bimodal tree for clustering, which successfully exploits the antipodal invariance of the coarse-to-high-res mapping of natural image patches and provides scalability to finer partitions of the underlying coarse patch space. During training an ensemble of such bimodal trees is computed, providing different linearizations of the mapping. The second and main contribution is a fast inference algorithm, which selects the most suitable mapping function within the tree ensemble for each patch by adopting a Local Naive Bayes formulation. The experimental validation shows promising scalability properties that reflect the suitability of the proposed model, which may also be generalized to other tasks. The resulting method is beyond one order of magnitude faster and performs objectively and subjectively better than the current state of the art.

****

POP Image Fusion - Derivative Domain Image Fusion Without Reintegration

Graham D. Finlayson, Alex E. Hayes; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 334-342

There are many applications where multiple images are fused to form a single summary greyscale or colour output, including computational photography (e.g. RGB-NIR), diffusion tensor imaging (medical), and remote sensing. Often, and intuitively, image fusion is carried out in the derivative domain. Here, a new composite fused derivative is found that best accounts for the detail across all images a nd then the resulting gradient field is reintegrated. However, the reintegration step generally hallucinates new detail (not appearing in any of the input image bands) including halo and bending artifacts. In this paper we avoid these hallucinated details by avoiding the reintegration step. Our work builds directly on the work of Socolinsky and Wolff who derive their equivalent gradient field from the per-pixel Di Zenzo structure tensor which is defined as the inner product of the image Jacobian. We show that the x- and y- derivatives of the projection of the original image onto the Principal characteristic vector of the Outer Product (POP) of the Jacobian generates the same equivalent gradient field. In so doing, we have derived a fused image that has the derivative structure we seek. Of course, this projection will be meaningful only where the Jacobian has non-zero derivatives, so we diffuse the projection directions using a bilateral filter before we calculate the fused image. The resulting POP fused image has maximal fused detail but avoids hallucinated artifacts. Experiments demonstrate our method delivers state of the art image fusion performance.

****

Adaptive Spatial-Spectral Dictionary Learning for Hyperspectral Image Denoising

Ying Fu, Antony Lam, Imari Sato, Yoichi Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 343-351

Hyperspectral imaging is beneficial in a diverse range of applications from diagnostic medicine, to agriculture, to surveillance to name a few. However, hyperspectral images often times suffer from degradation due to the limited light, which introduces noise into the imaging process. In this paper, we propose an effective model for hyperspectral image (HSI) denoising that considers underlying characteristics of HSIs: sparsity across the spatial-spectral domain, high correlation across spectra, and non-local self-similarity over space. We first exploit high correlation across spectra and non-local self-similarity over space in the noisy HSI to learn an adaptive spatial-spectral dictionary. Then, we employ the local and non-local sparsity of the HSI under the learned spatial-spectral dictionary to design an HSI denoising model, which can be effectively solved by an iterative numerical algorithm with parameters that are adaptively adjusted for different clusters and different noise levels. Experimental results on HSI denoising show that the proposed method can provide substantial improvements over the current state-of-the-art HSI denoising methods in terms of both objective metric and subjective visual quality.

****

Fully Connected Guided Image Filtering

Longquan Dai, Mengke Yuan, Feihu Zhang, Xiaopeng Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 352-360

This paper presents a linear time fully connected guided filter by introducing the minimum spanning tree (MST) to the guided filter (GF). Since the intensity based filtering kernel of GF is apt to overly smooth edges and the fixed-shape local box support region adopted by GF is not geometric-adaptive, our filter introduces an extra spatial term, the tree similarity, to the filtering kernel of GF and substitutes the box window with the implicit support region by establishing all-pairs-connections among pixels in the image and assigning the spatial-intensity-aware similarity to these connections. The adaptive implicit support region composed by the pixels with large kernel weights in the entire image domain has a big advantage over the predefined local box window in presenting the structure of an image for the reason that: 1, MST can efficiently present the structure of an image; 2, the kernel weight of our filter considers the tree distance defined on the MST. Due to these reasons, our filter achieves better edge-preserving results. We demonstrate the strength of the proposed filter in several applications. Experimental results show that our method produces better results than state-of-the-art methods.

****

Segment Graph Based Image Filtering: Fast Structure-Preserving Smoothing

Feihu Zhang, Longquan Dai, Shiming Xiang, Xiaopeng Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 361-369

In this paper, we design a new edge-aware structure, named segment graph, to rep resent the image and we further develop a novel double weighted average image filter (SGF) based on the segment graph. In our SGF, we use the tree distance on the segment graph to define the internal weight function of the filtering kernel, which enables the filter to smooth out high-contrast details and textures while preserving major image structures very well. While for the external weight function, we introduce a user specified smoothing window to balance the smoothing effects from each node of the segment graph. Moreover, we also set a threshold to adjust the edge-preserving performance. These advantages make the SGF more flexible in various applications and overcome the "halo" and "leak" problems appearing in most of the state-of-the-art approaches. Finally and importantly, we develop a linear algorithm for the implementation of our SGF, which has an O(N) time complexity for both gray-scale and high dimensional images, regardless of the kernel size and the intensity range. Typically, as one of the fastest edge-preserving filters, our CPU implementation achieves 0.15s per megapixel when performing filtering for 3-channel color images. The strength of the proposed filter is demonstrated by various applications, including stereo matching, optical flow, joint depth map upsampling, edge-preserving smoothing, edges detection, image abstraction and texture editing.

****

Deep Networks for Image Super-Resolution With Sparse Prior

Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, Thomas Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 370-378

Deep learning techniques have been successfully applied in many areas of computer vision, including low-level image restoration problems. For image super-resolution, several models based on deep neural networks have been recently proposed and attained superior performance that overshadows all previous handcrafted models. The question then arises whether large-capacity and data-driven models have become the dominant solution to the ill-posed super-resolution problem. In this paper, we argue that domain expertise represented by the conventional sparse coding model is still valuable, and it can be combined with the key ingredients of deep learning to achieve further improved results. We show that a sparse coding model particularly designed for super-resolution can be incarnated as a neural network, and trained in a cascaded structure from end to end. The interpretation of the network based on sparse coding leads to much more efficient and effective training, as well as a reduced model size. Our model is evaluated on a wide range of images, and shows clear advantage over existing state-of-the-art methods in terms of both restoration accuracy and human subjective quality.

****

Convolutional Color Constancy

Jonathan T. Barron; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 379-387

Color constancy is the problem of inferring the color of the light that illuminated a scene, usually so that the illumination color can be removed. Because this problem is underconstrained, it is often solved by modeling the statistical regularities of the colors of natural objects and illumination. In contrast, in this paper we reformulate the problem of color constancy as a 2D spatial localization task in a log-chrominance space, thereby allowing us to apply techniques from object detection and structured prediction to the color constancy problem. By directly learning how to discriminate between correctly white-balanced images and poorly white-balanced images, our model is able to improve performance on standard benchmarks by nearly 40%.

****

Learning Ordinal Relationships for Mid-Level Vision

Daniel Zoran, Phillip Isola, Dilip Krishnan, William T. Freeman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 388-396

We propose a framework that infers mid-level visual properties of an image by learning about ordinal relation- ships. Instead of estimating metric quantities directly, the system proposes pairwise relationship estimates for points in the input image. These sparse probabilistic ordinal mea- surements are globalized to create a dense output map of continuous metric measurements. Estimating order rel

a- tionships between pairs of points has several advantages over metric estimation: it solves a simpler problem than metric regression; humans are better at relative judgements, so data collection is easier; ordinal relationships are invari- ant to monotonic transformations of the data, thereby in- creasing the robustness of the system and providing qualitatively different information. We demonstrate that this frame- work works well on two important mid-level vision tasks: intrinsic image decomposition and depth from an RGB im- age. We train two systems with the same architecture on data from these two modalities. We provide an analysis of the resulting models, showing that they learn a number of simple rules to make ordinal decisions. We apply our algo-rithm to depth estimation, with good results, and intrinsic image decomposition, with state-of-the-art results.
****

Thin Structure Estimation With Curvature Regularization

Dmitrii Marin, Yuchen Zhong, Maria Drangova, Yuri Boykov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 397-405

Many applications in vision require estimation of thin structures such as boundary edges, surfaces, roads, blood vessels, neurons, etc. Unlike most previous approaches, we simultaneously detect and delineate thin structures with sub-pixel localization and real-valued orientation estimation. This is an ill-posed problem that requires regularization. We propose an objective function combining detection likelihoods with a prior minimizing curvature of the center-lines or surfaces. Unlike simple block-coordinate descent, we develop a novel algorithm that is able to perform joint optimization of location and detection variables more effectively. Our lower bound optimization algorithm applies to quadratic or absolute curvature. The proposed early vision framework is sufficiently general and it can be used in many higher-level applications. We illustrate the advantage of our approach on a range of 2D and 3D examples.
****

HARF: Hierarchy-Associated Rich Features for Salient Object Detection

Wenbin Zou, Nikos Komodakis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 406-414

The state-of-the-art salient object detection models are able to perform well for relatively simple scenes, yet for more complex ones, they still have difficulties in highlighting salient objects completely from background, largely due to the lack of sufficiently robust features for saliency prediction. To address such an issue, this paper proposes a novel hierarchy-associated feature construction framework for salient object detection, which is based on integrating elementary features from multi-level regions in a hierarchy. Furthermore, multi-layered deep learning features are introduced and incorporated as elementary features into this framework through a compact integration scheme. This leads to a rich feature representation, which is able to represent the context of the whole object/background and is much more discriminative as well as robust for salient object detection. Extensive experiments on the most widely used and challenging benchmark datasets demonstrate that the proposed approach substantially outperforms the state-of-the-art on salient object detection.
****

Deep Colorization

Zezhou Cheng, Qingxiong Yang, Bin Sheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 415-423

This paper investigates into the colorization problem which converts a grayscale image to a colorful version. This is a very difficult problem and normally requires manual adjustment to achieve artifact-free quality. For instance, it normally requires human-labelled color scribbles on the grayscale target image or a careful selection of colorful reference images (e.g., capturing the same scene in the grayscale target image). Unlike the previous methods, this paper aims at a high-quality fully-automatic colorization method. With the assumption of a perfect patch matching technique, the use of an extremely large-scale reference database (that contains sufficient color images) is the most reliable solution to the colorization problem. However, patch matching noise will increase with respect to the size of the reference database in practice. Inspired by the recent success

in deep learning techniques which provide amazing modeling of large-scale data, this paper re-formulates the colorization problem so that deep learning techniques can be directly employed. To ensure artifact-free quality, a joint bilateral filtering based post-processing step is proposed. Numerous experiments demonstrate that our method outperforms the state-of-art algorithms both in terms of quality and speed.
****

Image Matting With KL-Divergence Based Sparse Sampling
Levent Karacan, Aykut Erdem, Erkut Erdem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 424-432
Previous sampling-based image matting methods typically rely on certain heuristics in collecting representative samples from known regions, and thus their performance deteriorates if the underlying assumptions are not satisfied. To alleviate this, in this paper we take an entirely new approach and formulate sampling as a sparse subset selection problem where we propose to pick a small set of candidate samples that best explains the unknown pixels. Moreover, we describe a new distance measure for comparing two samples which is based on KL-divergence between the distributions of features extracted in the vicinity of the samples. Using a standard benchmark dataset for image matting, we demonstrate that our approach provides more accurate results compared with the state-of-the-art methods.
****

Intrinsic Decomposition of Image Sequences From Local Temporal Variations
Pierre-Yves Laffont, Jean-Charles Bazin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 433-441
We present a method for intrinsic image decomposition, which aims to decompose images into reflectance and shading layers. Our input is a sequence of images with varying illumination acquired by a static camera, e.g. an indoor scene with a moving light source or an outdoor timelapse. We leverage the local color variations observed over time to infer constraints on the reflectance and solve the ill-posed image decomposition problem. In particular, we derive an adaptive local energy from the observations of each local neighborhood over time, and integrate distant pairwise constraints to enforce coherent decomposition across all surfaces with consistent shading changes. Our method is solely based on multiple observations of a Lambertian scene under varying illumination and does not require user interaction, scene geometry, or an explicit lighting model. We compare our results with several intrinsic decomposition methods on a number of synthetic and captured datasets.
****

Low-Rank Tensor Approximation With Laplacian Scale Mixture Modeling for Multiframe Image Denoising
Weisheng Dong, Guangyu Li, Guangming Shi, Xin Li, Yi Ma; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 442-449
Patch-based low-rank models have shown effective in exploiting spatial redundancy of natural images especially for the application of image denoising. However, two-dimensional low-rank model can not fully exploit the spatio-temporal correlation in larger data sets such as multispectral images and 3D MRIs. In this work, we propose a novel low-rank tensor approximation framework with Laplacian Scale Mixture (LSM) modeling for multi-frame image denoising. First, similar 3D patches are grouped to form a tensor of d-order and high-order Singular Value Decomposition (HOSVD) is applied to the grouped tensor. Then the task of multiframe image denoising is formulated as a Maximum A Posterior (MAP) estimation problem with the LSM prior for tensor coefficients. Both unknown sparse coefficients and hidden LSM parameters can be efficiently estimated by the method of alternating optimization. Specifically, we have derived closed-form solutions for both subproblems. Experimental results on spectral and dynamic MRI images show that the proposed algorithm can better preserve the sharpness of important image structures and outperform several existing state-of-the-art multiframe denoising methods (e.g., BM4D and tensor dictionary learning).
****

Learning Parametric Distributions for Image Super-Resolution: Where Patch Matchi ng Meets Sparse Coding

Yongbo Li, Weisheng Dong, Guangming Shi, Xuemei Xie; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 450-458

Existing approaches toward Image super-resolution (SR) is often either data-driven (e.g., based on internet-scale matching and web image retrieval) or model-based (e.g., formulated as an Maximizing a Posterior estimation problem). The former is conceptually simple yet heuristic; while the latter is constrained by the fundamental limit of frequency aliasing. In this paper, we propose to develop a hybrid approach toward SR by combining those two lines of ideas. More specifically, the parameters underlying sparse distributions of desirable HR image patches are learned from a pair of LR image and retrieved HR images. Our hybrid approach can be interpreted as the first attempt of reconciling the difference between parametric and nonparametric models for low-level vision tasks. Experimental results show that the proposed hybrid SR method performs much better than existing state-of-the-art methods in terms of both subjective and objective image qualities.

****

Improving Image Restoration With Soft-Rounding

Xing Mei, Honggang Qi, Bao-Gang Hu, Siwei Lyu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 459-467

Several important classes of images such as text, barcode and pattern images have the property that pixels can only take a distinct subset of values. This knowledge can benefit the restoration of such images, but it has not been widely considered in current restoration methods. In this work, we describe an effective and efficient approach to incorporate the knowledge of distinct pixel values of the pristine images into the general regularized least squares restoration framework. We introduce a new regularizer that attains zero at the designated pixel values and becomes a quadratic penalty function in the intervals between them. When incorporated into the regularized least squares restoration framework, this regularizer leads to a simple and efficient step that resembles and extends the rounding operation, which we term as soft-rounding. We apply the soft-rounding enhanced solution to the restoration of binary text/barcode images and pattern images with multiple distinct pixel values. Experimental results show that soft-rounding enhanced restoration methods achieve significant improvement in both visual quality and quantitative measures (PSNR and SSIM). Furthermore, we show that this regularizer can also benefit the restoration of general natural images.

****

See the Difference: Direct Pre-Image Reconstruction and Pose Estimation by Differentiating HOG

Wei-Chen Chiu, Mario Fritz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 468-476

The Histogram of Oriented Gradient (HOG) descriptor has led to many advances in computer vision over the last decade and is still part of many state of the art approaches. We realize that the associated feature computation is piecewise differentiable and therefore many pipelines which build on HOG can be made differentiable. This lends to advanced introspection as well as opportunities for end-to-end optimization. We present our implementation of [?]HOG based on the auto-differentiation toolbox Chumpy and show applications to pre-image visualization and pose estimation which extends the existing differentiable renderer OpenDR pipeline. Both applications improve on the respective state-of-the-art HOG approaches.

****

An Efficient Statistical Method for Image Noise Level Estimation

Guangyong Chen, Fengyuan Zhu, Pheng Ann Heng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 477-485

In this paper, we address the problem of estimating noise level from a single image contaminated by additive zero-mean Gaussian noise. We first provide rigorous analysis on the statistical relationship between the noise variance and the eigenvalues of the covariance matrix of patches within an image, which shows that many state-of-the-art noise estimation methods underestimate the noise level of an image. To this end, we derive a new nonparametric algorithm for efficient nois e level estimation based on the observation that patches decomposed from a clean image often lie around a low-dimensional subspace. The performance of our method has been guaranteed both theoretically and empirically. Specifically, our method outperforms existing state-of-the-art algorithms on estimating noise level with the least executing time in our experiments. We further demonstrate that the denoising algorithm BM3D algorithm achieves optimal performance using noise variance estimated by our algorithm.

****

Contour Detection and Characterization for Asynchronous Event Sensors

Francisco Barranco, Ching L. Teo, Cornelia Fermuller, Yiannis Aloimonos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 486-494

The bio-inspired, asynchronous event-based dynamic vision sensor records temporal changes in the luminance of the scene at high temporal resolution. Since events are only triggered at significant luminance changes, most events occur at the boundary of objects and their parts. The detection of these contours is an essential step for further interpretation of the scene. This paper presents an approach to learn the location of contours and their border ownership using Structured Random Forests on event-based features that encode motion, timing, texture, and spatial orientations. The classifier integrates elegantly information over time by utilizing the classification results previously computed. Finally, the contour detection and boundary assignment are demonstrated in a layer-segmentation of the scene. Experimental results demonstrate good performance in boundary detection and segmentation.

****

Class-Specific Image Deblurring

Saeed Anwar, Cong Phuoc Huynh, Fatih Porikli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 495-503

In image deblurring, a fundamental problem is that the blur kernel suppresses a number of spatial frequencies that are difficult to recover reliably. In this paper, we explore the potential of a class-specific image prior for recovering spatial frequencies attenuated by the blurring process. Specifically, we devise a prior based on the class-specific subspace of image intensity responses to band-pass filters. We learn that the aggregation of these subspaces across all frequency bands serves as a good class-specific prior for the restoration of frequencies that cannot be recovered with generic image priors. In an extensive validation, our method, equipped with the above prior, yields greater image quality than many state-of-the-art methods by up to 5 dB in terms of image PSNR, across various image categories including portraits, cars, cats, pedestrians and household objects.

****

High-for-Low and Low-for-High: Efficient Boundary Detection From Deep Object Features and its Applications to High-Level Vision

Gedas Bertasius, Jianbo Shi, Lorenzo Torresani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 504-512

Most of the current boundary detection systems rely exclusively on low-level features, such as color and texture. However, perception studies suggest that humans employ object-level reasoning when judging if a particular pixel is a boundary. Inspired by this observation, in this work we show how to predict boundaries by exploiting object-level features from a pretrained object-classification network. Our method can be viewed as a "High-for-Low" approach where high-level object features inform the low-level boundary detection process. Our model achieves state-of-the-art performance on an established boundary detection benchmark and it is efficient to run. Additionally, we show that due to the semantic nature of our boundaries we can use them to aid a number of high-level vision tasks. We demonstrate that using our boundaries we improve the performance of state-of-the-art methods on the problems of semantic boundary labeling, semantic segmentation and object proposal generation. We can view this process as a "Low-for-High'" scheme, where low-level boundaries aid high-level vision tasks.   Thus, our contributions include a boundary detection system that is accurate, efficient, genera

lizes well to multiple datasets, and is also shown to improve existing state-of-the-art high-level vision methods on three distinct tasks.
****

Variational Depth Superresolution Using Example-Based Edge Representations
David Ferstl, Matthias Ruther, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 513-521

In this paper we propose a novel method for depth image superresolution which combines recent advances in example based upsampling with variational superresolution based on a known blur kernel. Most traditional depth superresolution approaches try to use additional high resolution intensity images as guidance for superresolution. In our method we learn a dictionary of edge priors from an external database of high and low resolution examples. In a novel variational sparse coding approach this dictionary is used to infer strong edge priors. Additionally to the traditional sparse coding constraints the difference in the overlap of neighboring edge patches is minimized in our optimization. These edge priors are used in a novel variational superresolution as anisotropic guidance of the higher order regularization. Both the sparse coding and the variational superresolution of the depth are solved based on a primal-dual formulation. In an exhaustive numerical and visual evaluation we show that our method clearly outperforms existing approaches on multiple real and synthetic datasets.
****

Conditioned Regression Models for Non-Blind Single Image Super-Resolution
Gernot Riegler, Samuel Schulter, Matthias Ruther, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 522-530

Single image super-resolution is an important task in the field of computer vision and finds many practical applications. Current state-of-the-art methods typically rely on machine learning algorithms to infer a mapping from low- to high-resolution images. These methods use a single fixed blur kernel during training and, consequently, assume the exact same kernel underlying the image formation process for all test images. However, this setting is not realistic for practical applications, because the blur is typically different for each test image. In this paper, we loosen this restrictive constraint and propose conditioned regression models (including convolutional neural networks and random forests) that can effectively exploit the additional kernel information during both, training and inference. This allows for training a single model, while previous methods need to be re-trained for every blur kernel individually to achieve good results, which we demonstrate in our evaluations. We also empirically show that the proposed conditioned regression models (i) can effectively handle scenarios where the blur kernel is different for each image and (ii) outperform related approaches trained for only a single kernel.
****

Video Super-Resolution via Deep Draft-Ensemble Learning
Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 531-539

We propose a new direction for fast video super-resolution (VideoSR) via a SR draft ensemble, which is defined as the set of high-resolution patch candidates before final image deconvolution. Our method contains two main components -- i.e., SR draft ensemble generation and its optimal reconstruction. The first component is to renovate traditional feedforward reconstruction pipeline and greatly enhance its ability to compute different super resolution results considering large motion variation and possible errors arising in this process. Then we combine SR drafts through the nonlinear process in a deep convolutional neural network (CNN). We analyze why this framework is proposed and explain its unique advantages compared to previous iterative methods to update different modules in passes. Promising experimental results are shown on natural video sequences.
****

Pan-Sharpening With a Hyper-Laplacian Penalty
Yiyong Jiang, Xinghao Ding, Delu Zeng, Yue Huang, John Paisley; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 540-548

Pan-sharpening is the task of fusing spectral information in low resolution mult ispectral images with spatial information in a corresponding high resolution panchromatic image. In such approaches, there is a trade-off between spectral and spatial quality, as well as computational efficiency. We present a method for pan-sharpening in which a sparsity-promoting objective function preserves both spatial and spectral content, and is efficient to optimize. Our objective incorporates the L1/2-norm in a way that can leverage recent computationally efficient methods, and L1 for which the alternating direction method of multipliers can be used. Additionally, our objective penalizes image gradients to enforce high resolution fidelity, and exploits the Fourier domain for further computational efficiency. Visual quality metrics demonstrate that our proposed objective function can achieve higher spatial and spectral resolution than several previous well-known methods with competitive computational efficiency.

****

Video Restoration Against Yin-Yang Phasing

Xiaolin Wu, Zhenhao Li, Xiaowei Deng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 549-557

A common video degradation problem, which is largely untreated in literature, is what we call Yin-Yang Phasing (YYP). YYP is characterized by involuntary, dramatic flip-flop in the intensity and possibly chromaticity of an object as the video plays. Such temporal artifacts occur under ill illumination conditions and are triggered by object or/and camera motions, which mislead the settings of camera's auto-exposure and white point. In this paper, we investigate the problem and propose a video restoration technique to suppress YYP artifacts and retain temporal consistency of objects appearance via inter-frame, spatially-adaptive, optimal tone mapping. The video quality can be further improved by a novel image enhancer designed in Weber's perception principle and by exploiting the second-order statistics of the scene. Experimental results are encouraging, pointing to an effective, practical solution for a common but surprisingly understudied problem.

****

Rolling Shutter Super-Resolution

Abhijith Punnappurath, Vijay Rengarajan, A.N. Rajagopalan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 558-566

Classical multi-image super-resolution (SR) algorithms, designed for CCD cameras, assume that the motion among the images is global. But CMOS sensors that have increasingly started to replace their more expensive CCD counterparts in many applications do not respect this assumption if there is a motion of the camera relative to the scene during the exposure duration of an image because of the row-wise acquisition mechanism. In this paper, we study the hitherto unexplored topic of multi-image SR in CMOS cameras. We initially develop an SR observation model that accounts for the row-wise distortions called the ``rolling shutter'' (RS) effect observed in images captured using non-stationary CMOS cameras. We then propose a unified RS-SR framework to obtain an RS-free high-resolution image (and the row-wise motion) from distorted low-resolution images. We demonstrate the efficacy of the proposed scheme using synthetic data as well as real images captured using a hand-held CMOS camera. Quantitative and qualitative assessments reveal that our method significantly advances the state-of-the-art.

****

Learning Large-Scale Automatic Image Colorization

Aditya Deshpande, Jason Rock, David Forsyth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 567-575

We describe an automated method for image colorization that learns to colorize from examples. Our method exploits a LEARCH framework to train a quadratic objective function in the chromaticity maps, comparable to a Gaussian random field. The coefficients of the objective function are conditioned on image features, using a random forest. The objective function admits correlations on long spatial scales, and can control spatial error in the colorization of the image. Images are then colorized by minimizing this objective function. We demonstrate that our method strongly outperforms a natural baseline on large-scale experiments with images of real scenes using a demanding loss function. We demonstrate that l earning a model that is conditioned on scene produces improved results. We show how to incorporate a desired color histogram into the objective function, and that doing so can lead to further improvements in results.

****

Compression Artifacts Reduction by a Deep Convolutional Network

Chao Dong, Yubin Deng, Chen Change Loy, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 576-584

Lossy compression introduces complex compression artifacts, particularly the blocking artifacts, ringing effects and blurring. Existing algorithms either focus on removing blocking artifacts and produce blurred output, or restores sharpened images that are accompanied with ringing effects. Inspired by the deep convolutional networks (DCN) on super-resolution, we formulate a compact and efficient network for seamless attenuation of different compression artifacts. We also demonstrate that a deeper model can be effectively trained with the features learned in a shallow network. Following a similar "easy to hard" idea, we systematically investigate several practical transfer settings and show the effectiveness of transfer learning in low level vision problems. Our method shows superior performance than the state-of-the-arts both on the benchmark datasets and the real-world use cases (i.e. Twitter).

****

Multiple-Hypothesis Affine Region Estimation With Anisotropic LoG Filters

Takahiro Hasegawa, Mitsuru Ambai, Kohta Ishikawa, Gou Koutaki, Yuji Yamauchi, Takayoshi Yamashita, Hironobu Fujiyoshi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 585-593

We propose a method for estimating multiple-hypothesis affine regions from a keypoint by using an anisotropic Laplacian-of-Gaussian (LoG) filter. Although conventional affine region detectors, such as Hessian/Harris-Affine, iterate to find an affine region that fits a given image patch, such iterative searching is adversely affected by an initial point. To avoid this problem, we allow multiple detections from a single keypoint. We demonstrate that the responses of all possible anisotropic LoG filters can be efficiently computed by factorizing them in a similar manner to spectral SIFT. A large number of LoG filters that are densely sampled in a parameter space are reconstructed by a weighted combination of a limited number of representative filters, called ``eigenfilters", by using singular value decomposition. Also, the reconstructed filter responses of the sampled parameters can be interpolated to a continuous representation by using a series of proper functions. This results in efficient multiple extrema searching in a continuous space. Experiments revealed that our method has higher repeatability than the conventional methods.

****

A Self-Paced Multiple-Instance Learning Framework for Co-Saliency Detection

Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, Junwei Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 594-602

As an interesting and emerging topic, co-saliency detection aims at simultaneously extracting common salient objects in a group of images. Traditional co-saliency detection approaches rely heavily on human knowledge for designing hand-crafted metrics to explore the intrinsic patterns underlying co-salient objects. Such strategies, however, always suffer from poor generalization capability to flexibly adapt various scenarios in real applications, especially due to their lack of insightful understanding of the biological mechanisms of human visual co-attention. To alleviate this problem, we propose a novel framework for this task, by naturally reformulating it as a multiple-instance learning (MIL) problem and further integrating it into a self-paced learning (SPL) regime. The proposed framework on one hand is capable of fitting insightful metric measurements and discovering common patterns under co-salient regions in a self-learning way by MIL, and on the other hand tends to promise the learning reliability and stability by simulating the human learning process through SPL. Experiments on benchmark datasets have demonstrated the effectiveness of the proposed framework as compared with the state-of-the-arts.

****

External Patch Prior Guided Internal Clustering for Image Denoising

Fei Chen, Lei Zhang, Huimin Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 603-611

Natural image modeling plays a key role in many vision problems such as image denoising. Image priors are widely used to regularize the denoising process, which is an illposed inverse problem. One category of denoising methods exploit the priors (e.g., TV, sparsity) learned from external clean images to reconstruct the given noisy image, while another category of methods exploit the internal prior (e.g., self-similarity) to reconstruct the latent image. Though the internal prior based methods have achieved impressive denoising results, the improvement of visual quality will become very difficult with the increase of noise level. In this paper, we propose to exploit image external patch prior and internal self-similarity prior jointly, and develop an external patch prior guided internal clustering algorithm for image denoising. It is known that natural image patches form multiple subspaces. By utilizing Gaussian mixture models (GMMs) learning, image similar patches can be clustered and the subspaces can be learned. The learned GMMs from clean images are then used to guide the clustering of noisypatches of the input noisy images, followed by a low-rank approximation process to estimate the latent subspace for image recovery. Numerical experiments show that the proposed method outperforms many state-of-the-art denoising algorithms such as BM3D and WNNM.

****

Self-Calibration of Optical Lenses

Michael Hirsch, Bernhard Scholkopf; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 612-620

Even high-quality lenses suffer from optical aberrations, especially when used at full aperture. Furthermore, there are significant lens-to-lens deviations due to manufacturing tolerances, often rendering current software solutions like DxO, Lightroom, and PTLens insufficient as they don't adapt and only include generic lens blur models. We propose a method that enables the self-calibration of lenses from a natural image, or a set of such images. To this end we develop a machine learning framework that is able to exploit several recorded images and distills the available information into an accurate model of the considered lens.

****

Illumination Robust Color Naming via Label Propagation

Yuanliu liu, Zejian Yuan, Badong Chen, Jianru Xue, Nanning Zheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 621-629

Color composition is an important property for many computer vision tasks like image retrieval and object classification. In this paper we address the problem of inferring the color composition of the intrinsic reflectance of objects, where the shadows and highlights may change the observed color dramatically. We achieve this through color label propagation without recovering the intrinsic reflectance beforehand. Specifically, the color labels are propagated between regions sharing the same reflectance, and the direction of propagation is promoted to be from regions under full illumination and normal view angles to abnormal regions. We detect shadowed and highlighted regions as well as pairs of regions that have similar reflectance. A joint inference process is adopted to trim the inconsistent identities and connections. For evaluation we collect three datasets of images under noticeable highlights and shadows. Experimental results show that our model can effectively describe the color composition of real-world images.

****

Unsupervised Cross-Modal Synthesis of Subject-Specific Scans

Raviteja Vemulapalli, Hien Van Nguyen, Shaohua Kevin Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 630-638

Recently, cross-modal synthesis of subject-specific scans has been receiving significant attention from the medical imaging community. Though various synthesis approaches have been introduced in the recent past, most of them are either tailored to a specific application or proposed for the supervised setting, i.e., they assume the availability of training data from the same set of subjects in both source and target modalities. But, collecting multiple scans from each subject is undesirable. Hence, to address this issue, we propose a general unsupervised cross-modal medical image synthesis approach that works without paired training data. Given a source modality image of a subject, we first generate multiple target modality candidate values for each voxel independently using cross-modal nearest neighbor search. Then, we select the best candidate values jointly for all the voxels by simultaneously maximizing a global mutual information cost function and a local spatial consistency cost function. Finally, we use coupled sparse representation for further refinement of synthesized images. Our experiments on generating T1-MRI brain scans from T2-MRI and vice versa demonstrate that the synthesis capability of the proposed unsupervised approach is comparable to various state-of-the-art supervised approaches in the literature.

****

Learning to Boost Filamentary Structure Segmentation

Lin Gu, Li Cheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 639-647

The challenging problem of filamentary structure segmentation has a broad range of applications in biological and medical fields. A critical yet challenging issue remains on how to detect and restore the small filamentary fragments from backgrounds: The small fragments are of diverse shapes and appearances, meanwhile the backgrounds could be cluttered and ambiguous. Focusing on this issue, this paper proposes an iterative two-step learning-based approach to boost the performance based on a base segmenter arbitrarily chosen from a number of existing segmenters: We start with an initial partial segmentation where the filamentary structure obtained is of high confidence based on this existing segmenter. We also define a scanning horizon as epsilon balls centred around the partial segmentation result. Step one of our approach centers on a data-driven latent classification tree model to detect the filamentary fragments. This model is learned via a training process, where a large number of distinct local figure/background separation scenarios are established and geometrically organized into a tree structure. Step two spatially restores the isolated fragments back to the current partial segmentation, which is accomplished by means of completion fields and matting. Both steps are then alternated with the growth of partial segmentation result, until the input image space is entirely explored. Our approach is rather generic and can be easily augmented to a wide range of existing supervised/unsupervised segmenters to produce an improved result. This has been empirically verified on specific filamentary structure segmentation tasks: retinal blood vessel segmentation as well as neuronal segmentations, where noticeable improvement has been shown over the original state-of-the-arts.

****

Weakly-Supervised Structured Output Learning With Flexible and Latent Graphs Using High-Order Loss Functions

Gustavo Carneiro, Tingying Peng, Christine Bayer, Nassir Navab; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 648-656

We introduce two new structured output models that use a latent graph, which is flexible in terms of the number of nodes and structure, where the training process minimises a high-order loss function using a weakly annotated training set. These models are developed in the context of microscopy imaging of malignant tumours, where the estimation of the number and proportion of classes of microcirculatory supply units (MCSU) is important in the assessment of the efficacy of common cancer treatments (an MCSU is a region of the tumour tissue supplied by a microvessel). The proposed methodologies take as input multimodal microscopy images of a tumour, and estimate the number and proportion of MCSU classes. This estimation is facilitated by the use of an underlying latent graph (not present in the manual annotations), where each MCSU is represented by a node in this graph, labelled with the MCSU class and image location. The training process uses the manual weak annotations available, consisting of the number of MCSU classes per training image, where the training objective is the minimisation of a high-order loss function based on the norm of the error between the manual and estimated annotations. One of the models proposed is based on a new flexible latent struct ure support vector machine (FLSSVM) and the other is based on a deep convolutional neural network (DCNN) model. Using a dataset of 89 weakly annotated pairs of multimodal images from eight tumours, we show that the quantitative results from DCNN are superior, but the qualitative results from FLSSVM are better and both display high correlation values regarding the number and proportion of MCSU classes compared to the manual annotations.

****

Efficient Classifier Training to Minimize False Merges in Electron Microscopy Segmentation

Toufiq Parag, Dan C. Ciresan, Alessandro Giusti; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 657-665

The prospect of neural reconstruction from Electron Microscopy (EM) images has been elucidated by the automatic segmentation algorithms. Although segmentation algorithms eliminate the necessity of tracing the neurons by hand, significant manual effort is still essential for correcting the mistakes they make. A considerable amount of human labor is also required for annotating groundtruth volumes for training the classifiers of a segmentation framework. It is critically important to diminish the dependence on human interaction in the overall reconstruction system. This study proposes a novel classifier training algorithm for EM segmentation aimed to reduce the amount of manual effort demanded by the groundtruth annotation and error refinement tasks. Instead of using an exhaustive pixel level groundtruth, an active learning algorithm is proposed for sparse labeling of pixel and boundaries of superpixels. Because over-segmentation errors are in general more tolerable and easier to correct than the under-segmentation errors, our algorithm is designed to prioritize minimization of false-merges over false-split mistakes. Our experiments on both 2D and 3D data suggest that the proposed method yields segmentation outputs that are more amenable to neural reconstruction than those of existing methods.

****

On Statistical Analysis of Neuroimages With Imperfect Registration

Won Hwa Kim, Sathya N. Ravi, Sterling C. Johnson, Ozioma C. Okonkwo, Vikas Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 666-674

A variety of studies in neuroscience/neuroimaging seek to perform statistical inference on the acquired brain image scans for diagnosis as well as understanding the pathological manifestation of diseases. To do so, an important first step is to register (or co-register) all of the image data into a common coordinate system. This permits meaningful comparison of the intensities at each voxel across groups (e.g., diseased versus healthy) to evaluate the effects of the disease and/or use machine learning algorithms in a subsequent step. But errors in the underlying registration make this problematic, they either decrease the statistical power or make the follow-up inference tasks less effective/accurate. In this paper, we derive a novel algorithm which offers immunity to local errors in the underlying deformation field obtained from registration procedures. By deriving a deformation invariant representation of the image, the downstream analysis can be made more robust as if one had access to a (hypothetical) far superior registration procedure. Our algorithm is based on recent work on Scattering coefficients. Using this as a starting point, we show how results from harmonic analysis (especially, non-Euclidean wavelets) yields strategies for designing deformation and additive noise invariant representations of large 3-D brain image volumes. We present a set of results on synthetic and real brain images where we achieve robust statistical analysis even in the presence of substantial deformation errors; here, standard analysis procedures significantly under-perform and fail to identify the true signal.

****

Convex Optimization With Abstract Linear Operators

Steven Diamond, Stephen Boyd; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 675-683

We introduce a convex optimization modeling framework that transforms a convex optimization problem expressed in a form natural and convenient for the user into an equivalent cone program in a way that preserves fast linear transforms in the original problem. By representing linear functions in the transformation process not as matrices, but as graphs that encode composition of abstract linear operators, we arrive at a matrix-free cone program, i.e., one whose data matrix is represented by an abstract linear operator and its adjoint. This cone program can then be solved by a matrix-free cone solver. By combining the matrix-free modeling framework and cone solver, we obtain a general method for efficiently solving convex optimization problems involving fast linear transforms.

****

Building Dynamic Cloud Maps From the Ground Up

Calvin Murdock, Nathan Jacobs, Robert Pless; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 684-692

Satellite imagery of cloud cover is extremely important for understanding and predicting weather. We demonstrate how this imagery can be constructed "from the ground up" without requiring expensive geo-stationary satellites. This is accomplished through a novel approach to approximate continental-scale cloud maps using only ground-level imagery from publicly-available webcams. We collected a year's worth of satellite data and simultaneously-captured, geo-located outdoor webcam images from 4388 sparsely distributed cameras across the continental USA. The satellite data is used to train a dynamic model of cloud motion alongside 4388 regression models (one for each camera) to relate ground-level webcam data to the satellite data at the camera's location. This novel application of large-scale computer vision to meteorology and remote sensing is enabled by a smoothed, hierarchically-regularized dynamic texture model whose system dynamics are driven to remain consistent with measurements from the geo-located webcams. We show that our hierarchical model is better able to incorporate sparse webcam measurements resulting in more accurate cloud maps in comparison to a standard dynamic textures implementation. Finally, we demonstrate that our model can be successfully applied to other natural image sequences from the DynTex database, suggesting a broader applicability of our method.

****

A Versatile Learning-Based 3D Temporal Tracker: Scalable, Robust, Online

David Joseph Tan, Federico Tombari, Slobodan Ilic, Nassir Navab; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 693-701

This paper proposes a temporal tracking algorithm based on Random Forest that uses depth images to estimate and track the 3D pose of a rigid object in real-time. Compared to the state of the art aimed at the same goal, our algorithm holds important attributes such as high robustness against holes and occlusion, low computational cost of both learning and tracking stages, and low memory consumption. These are obtained (a) by a novel formulation of the learning strategy, based on a dense sampling of the camera viewpoints and learning independent trees from a single image for each camera view; as well as, (b) by an insightful occlusion handling strategy that enforces the forest to recognize the object's local and global structures. Due to these attributes, we report state-of-the-art tracking accuracy on benchmark datasets, and accomplish remarkable scalability with the number of targets, being able to simultaneously track the pose of over a hundred objects at 30 fps with an off-the-shelf CPU. In addition, the fast learning time enables us to extend our algorithm as a robust online tracker for model-free 3D objects under different viewpoints and appearance changes as demonstrated by the experiments.

****

Realtime Edge-Based Visual Odometry for a Monocular Camera

Juan Jose Tarrio, Sol Pedre; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 702-710

In this work we present a novel algorithm for realtime visual odometry for a monocular camera. The main idea is to develop an approach between classical feature-based visual odometry systems and modern direct dense/semi-dense methods, trying to benefit from the best attributes of both. Similar to feature-based systems, we extract information from the images, instead of working with raw image intensities as direct methods. In particular, the information extracted are the edges present in the image, while the rest of the algorithm is designed to take advantage of the structural information provided when pixels are treated as edges. Edge extraction is an efficient and higly parallelizable operation. The edge depth information extracted is dense enough to allow acceptable surface fitting, similar to modern semi-dense methods. This is a valuable attribute that feature-based odometry lacks. Experimental results show that the proposed method has similar drift than state of the art feature-based and direct methods, and is a simple algorithm that runs at realtime and can be parallelized. Finally, we have also developed an inertial aided version that successfully stabilizes an unmanned air vehicle in complex indoor environments using only a frontal camera, while running the complete solution in the embedded hardware on board the vehicle.

****

Fill and Transfer: A Simple Physics-Based Approach for Containability Reasoning

Lap-Fai Yu, Noah Duncan, Sai-Kit Yeung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 711-719

The visual perception of object affordances has emerged as a useful ingredient for building powerful computer vision and robotic applications. In this paper we introduce a novel approach to reason about liquid containability - the affordance of containing liquid. Our approach analyzes container objects based on two simple physical processes: the Fill and Transfer of liquid. First, it reasons about whether a given 3D object is a liquid container and its best filling direction. Second, it proposes directions to transfer its contained liquid to the outside while avoiding spillage. We compare our simplified model with a common fluid dynamics simulation and demonstrate that our algorithm makes human-like choices about the best directions to fill containers and transfer liquid from them. We apply our approach to reason about the containability of several real-world objects acquired using a consumer-grade depth camera.

****

On Linear Structure From Motion for Light Field Cameras

Ole Johannsen, Antonin Sulc, Bastian Goldluecke; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 720-728

We present a novel approach to relative pose estimation which is tailored to 4D light field cameras. From the relationships between scene geometry and light field structure and an analysis of the light field projection in terms of Pluecker ray coordinates, we deduce a set of linear constraints on ray space correspondences between a light field camera pair. These can be applied to infer relative pose of the light field cameras and thus obtain a point cloud reconstruction of the scene. While the proposed method has interesting relationships to pose estimation for generalized cameras based on ray-to-ray correspondence, our experiments demonstrate that our approach is both more accurate and computationally more efficient. It also compares favourably to direct linear pose estimation based on aligning the 3D point clouds obtained by reconstructing depth for each individual light field. To further validate the method, we employ the pose estimates to merge light fields captured with hand-held consumer light field cameras into refocusable panoramas.

****

3D Object Reconstruction From Hand-Object Interactions

Dimitrios Tzionas, Juergen Gall; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 729-737

Recent advances have enabled 3d object reconstruction approaches using a single off-the-shelf RGB-D camera. Although these approaches are successful for a wide range of object classes, they rely on stable and distinctive geometric or texture features. Many objects like mechanical parts, toys, household or decorative articles, however, are textureless and characterized by minimalistic shapes that are simple and symmetric. Existing in-hand scanning systems and 3d reconstruction techniques fail for such symmetric objects in the absence of highly distinctive features. In this work, we show that extracting 3d hand motion for in-hand scanning effectively facilitates the reconstruction of even featureless and highly symmetric objects and we present an approach that fuses the rich additional information of hands into a 3d reconstruction pipeline, significantly contributing to the state-of-the-art of in-hand scanning.

****

Minimal Solvers for 3D Geometry From Satellite Imagery
Enliang Zheng, Ke Wang, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 738-746

We propose two novel minimal solvers which advance the state of the art in satellite imagery processing. Our methods are efficient and do not rely on the prior existence of complex inverse mapping functions to correlate 2D image coordinates and 3D terrain. Our first solver improves on the stereo correspondence problem for satellite imagery, in that we provide an exact image-to-object space mapping (where prior methods were inaccurate). Our second solver provides a novel mechanism for 3D point triangulation, which has improved robustness and accuracy over prior techniques. Given the usefulness and ubiquity of satellite imagery, our proposed methods allow for improved results in a variety of existing and future applications.

****

An Efficient Minimal Solution for Multi-Camera Motion
Jonathan Ventura, Clemens Arth, Vincent Lepetit; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 747-755

We propose an efficient method for estimating the motion of a multi-camera rig from a minimal set of feature correspondences. Existing methods for solving the multi-camera relative pose problem require extra correspondences, are slow to compute, and/or produce a multitude of solutions. Our solution uses a first-order approximation to relative pose in order to simplify the problem and produce an accurate estimate quickly. The solver is applicable to sequential multi-camera motion estimation and is fast enough for real-time implementation in a random sampling framework. Our experiments show that our approach is both stable and efficient on challenging test sequences.

****

Learning Shape, Motion and Elastic Models in Force Space
Antonio Agudo, Francesc Moreno-Noguer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 756-764

In this paper, we address the problem of simultaneously recovering the 3D shape and pose of a deformable and potentially elastic object from 2D motion. This is a highly ambiguous problem typically tackled by using low-rank shape and trajectory constraints. We show that formulating the problem in terms of a low-rank force space that induces the deformation, allows for a better physical interpretation of the resulting priors and a more accurate representation of the actual object's behavior. However, this comes at the price of, besides force and pose, having to estimate the elastic model of the object. For this, we use an Expectation Maximization strategy, where each of these parameters are successively learned within partial M-steps, while robustly dealing with missing observations. We thoroughly validate the approach on both mocap and real sequences, showing more accurate 3D reconstructions than state-of-the-art, and additionally providing an estimate of the full elastic model with no a priori information.

****

A Versatile Scene Model With Differentiable Visibility Applied to Generative Pose Estimation
Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 765-773

Generative reconstruction methods compute the 3D configuration (such as pose and/or geometry) of a shape by optimizing the overlap of the projected 3D shape model with images. Proper handling of occlusions is a big challenge, since the visibility function that indicates if a surface point is seen from a camera can often not be formulated in closed form, and is in general discrete and non-differentiable at occlusion boundaries. We present a new scene representation that enables an analytically differentiable closed-form formulation of surface visibility. In contrast to previous methods, this yields smooth, analytically differentiable, and efficient to optimize pose similarity energies with rigorous occlusion han

dling, fewer local minima, and experimentally verified improved convergence of numerical optimization. The underlying idea is a new image formation model that represents opaque objects by a translucent medium with a smooth Gaussian density distribution which turns visibility into a smooth phenomenon. We demonstrate the advantages of our versatile scene model in several generative pose estimation problems, namely marker-less multi-object pose estimation, marker-less human motion capture with few cameras, and image-based 3D geometry estimation.
****

Semantic Pose Using Deep Networks Trained on Synthetic RGB-D
Jeremie Papon, Markus Schoeler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 774-782

In this work we address the problem of indoor scene understanding from RGB-D images. Specifically, we propose to find instances of common furniture classes, their spatial extent, and their pose with respect to generalized class models. To accomplish this, we use a deep, wide, multi-output convolutional neural network (CNN) that predicts class, pose, and location of possible objects simultaneously. To overcome the lack of large annotated RGB-D training sets (especially those with pose), we use an on-the-fly rendering pipeline that generates realistic cluttered room scenes in parallel to training. We then perform transfer learning on the relatively small amount of publicly available annotated RGB-D data, and find that our model is able to successfully annotate even highly challenging real scenes. Importantly, our trained network is able to understand noisy and sparse observations of highly cluttered scenes with a remarkable degree of accuracy, inferring class and pose from a very limited set of cues. Additionally, our neural network is only moderately deep and computes class, pose and position in tandem, so the overall run-time is significantly faster than existing methods, estimating all output parameters simultaneously in parallel.
****

Exploiting High Level Scene Cues in Stereo Reconstruction
Simon Hadfield, Richard Bowden; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 783-791

We present a novel approach to 3D reconstruction which is inspired by the human visual system. This system unifies standard appearance matching and triangulation techniques with higher level reasoning and scene understanding, in order to resolve ambiguities between different interpretations of the scene. The types of reasoning integrated in the approach includes recognising common configurations of surface normals and semantic edges (e.g. convex, concave and occlusion boundaries). We also recognise the coplanar, collinear and symmetric structures which are especially common in man made environments.
****

Point Triangulation Through Polyhedron Collapse Using the l[?] Norm
Simon Donne, Bart Goossens, Wilfried Philips; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 792-800

Multi-camera triangulation of feature points based on a minimisation of the overall L2 reprojection error can get stuck in suboptimal local minima or require slow global optimisation. For this reason, researchers have proposed optimising the L-infinity norm of the L2 single view reprojection errors, which avoids the problem of local minima entirely. In this paper we present a novel method for L-infinity triangulation that minimizes the L-infinity norm of the L-infinity reprojection errors: this apparently small difference leads to a much faster but equally accurate solution which is related to the MLE under the assumption of uniform noise. The proposed method adopts a new optimisation strategy based on solving simple quadratic equations. This stands in contrast with the fastest existing methods, which solve a sequence of more complex auxiliary Linear Programming or Second Order Cone Problems. The proposed algorithm performs well: for triangulation, it achieves the same accuracy as existing techniques while executing faster and being straightforward to implement.
****

Optimizing the Viewing Graph for Structure-From-Motion
Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, Marc Pollefeys; P

roceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 801-809

The viewing graph represents a set of views that are related by pairwise relative geometries. In the context of Structure-from-Motion (SfM), the viewing graph is the input to the incremental or global estimation pipeline. Much effort has been put towards developing robust algorithms to overcome potentially inaccurate relative geometries in the viewing graph during SfM. In this paper, we take a fundamentally different approach to SfM and instead focus on improving the quality of the viewing graph before applying SfM. Our main contribution is a novel optimization that improves the quality of the relative geometries in the viewing graph by enforcing loop consistency constraints with the epipolar point transfer. We show that this optimization greatly improves the accuracy of relative poses in the viewing graph and removes the need for filtering steps or robust algorithms typically used in global SfM methods. In addition, the optimized viewing graph can be used to efficiently calibrate cameras at scale. We combine our viewing graph optimization and focal length calibration into a global SfM pipeline that is more efficient than existing approaches. To our knowledge, ours is the first global SfM pipeline capable of handling uncalibrated image sets.

****

Intrinsic Scene Decomposition From RGB-D images

Mohammed Hachama, Bernard Ghanem, Peter Wonka; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 810-818

In this paper, we address the problem of computing an intrinsic decomposition of the colors of a surface into an albedo and a shading term. The surface is reconstructed from a single or multiple RGB-D images of a static scene obtained from different views. We thereby extend and improve existing works in the area of intrinsic image decomposition. In a variational framework, we formulate the problem as a minimization of an energy composed of two terms: a data term and a regularity term. The first term is related to the image formation process and expresses the relation between the albedo, the surface normals, and the incident illumination. We use an affine shading model, a combination of a Lambertian model, and an ambient lighting term. This model is relevant for Lambertian surfaces. When available, multiple views can be used to handle view-dependent non-Lambertian reflections. The second term contains an efficient combination of l2 and l1-regularizers on the illumination vector field and albedo respectively. Unlike most previous approaches, especially Retinex-like techniques, these terms do not depend on the image gradient or texture, thus reducing the mixing shading/reflectance artifacts and leading to better results. The obtained non-linear optimization problem is efficiently solved using a cyclic block coordinate descent algorithm. Our method outperforms a range of state-of-the-art algorithms on a popular benchmark dataset.

****

3D Hand Pose Estimation Using Randomized Decision Forest With Segmentation Index Points

Peiyi Li, Haibin Ling, Xi Li, Chunyuan Liao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 819-827

In this paper, we propose a real-time 3D hand pose estimation algorithm using the randomized decision forest framework. Our algorithm takes a depth image as input and generates a set of skeletal joints as output. Previous decision forest-based methods often give labels to all points in a point cloud at a very early stage and vote for the joint locations. By contrast, our algorithm only tracks a set of more flexible virtual landmark points, named segmentation index points (SIPs), before reaching the final decision at a leaf node. Roughly speaking, a SIP represents the centroid of a subset of skeletal joints, which are to be located at the leaves of the branch expanded from the SIP. Inspired by recent latent regression forest-based hand pose estimation framework (Tang et al. 2014), we integrate SIP into the framework with several important improvements: First, we devise a new forest growing strategy, whose decision is made using a randomized feature guided by SIPs. Second, we speed-up the training procedure since only SIPs, not the skeletal joints, are estimated at non-leaf nodes. Third, the experimental results on public benchmark datasets show clearly the advantage of the proposed algorithm over previous state-of-the-art methods, and our algorithm runs at 55.5 fps on a normal CPU without parallelism.

****

Accurate Camera Calibration Robust to Defocus Using a Smartphone

Hyowon Ha, Yunsu Bok, Kyungdon Joo, Jiyoung Jung, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 828-836

We propose a novel camera calibration method for defocused images using a smartphone under the assumption that the defocus blur is modeled as a convolution of a sharp image with a Gaussian point spread function (PSF). In contrast to existing calibration approaches which require well-focused images, the proposed method achieves accurate camera calibration with severely defocused images. This robustness to defocus is due to the proposed set of unidirectional binary patterns, which simplifies 2D Gaussian deconvolution to a 1D Gaussian deconvolution problem with multiple observations. By capturing the set of patterns consecutively displayed on a smartphone, we formulate the feature extraction as a deconvolution problem to estimate feature point locations in sub-pixel accuracy and the blur kernel in each location. We also compensate the error in camera parameters due to refraction of the glass panel of the display device. We evaluate the performance of the proposed method on synthetic and real data. Even under severe defocus, our method shows accurate camera calibration result.

****

High Quality Structure From Small Motion for Rolling Shutter Cameras

Sunghoon Im, Hyowon Ha, Gyeongmin Choe, Hae-Gon Jeon, Kyungdon Joo, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 837-845

We present a practical 3D reconstruction method to obtain a high-quality dense depth map from narrow-baseline image sequences captured by commercial digital cameras, such as DSLRs or mobile phones. Depth estimation from small motion has gained interest as a means of various photographic editing, but important limitations present themselves in the form of depth uncertainty due to a narrow baseline and rolling shutter. To address these problems, we introduce a novel 3D reconstruction method from narrow-baseline image sequences that effectively handles the effects of a rolling shutter that occur from most of commercial digital cameras. Additionally, we present a depth propagation method to fill in the holes associated with the unknown pixels based on our novel geometric guidance model. Both qualitative and quantitative experimental results show that our new algorithm consistently generates better 3D depth maps than those by the state-of-the-art method.

****

Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction

Paulo F. U. Gotardo, Tomas Simon, Yaser Sheikh, Iain Matthews; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 846-854

Photometric stereo (PS) is an established technique for high-detail reconstruction of 3D geometry and appearance. To correct for surface integration errors, PS is often combined with multiview stereo (MVS). With dynamic objects, PS reconstruction also faces the problem of computing optical flow (OF) for image alignment under rapid changes in illumination. Current PS methods typically compute optical flow and MVS as independent stages, each one with its own limitations and errors introduced by early regularization. In contrast, scene flow methods estimate geometry and motion, but lack the fine detail from PS. This paper proposes photogeometric scene flow (PGSF) for high-quality dynamic 3D reconstruction. PGSF performs PS, OF, and MVS simultaneously. It is based on two key observations: (i) while image alignment improves PS, PS allows for surfaces to be relit to improve alignment; (ii) PS provides surface gradients that render the smoothness term in MVS unnecessary, leading to truly data-driven, continuous depth estimates. This synergy is demonstrated in the quality of the resulting RGB appearance, 3D geometry, and 3D motion.

****

Blur-Aware Disparity Estimation From Defocus Stereo Images

Ching-Hui Chen, Hui Zhou, Timo Ahonen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 855-863

Defocus blur usually causes performance degradation in establishing the visual correspondence between stereo images. We propose a blur-aware disparity estimation method that is robust to the mismatch of focus in stereo images. The relative blur resulting from the mismatch of focus between stereo images is approximated as the difference of the square diameters of the blur kernels. Based on the defocus and stereo model, we propose the relative blur versus disparity (RBD) model that characterizes the relative blur as a second-order polynomial function of disparity. Our method alternates between RBD model update and disparity update in each iteration. The RBD model in return refines the disparity estimation by updating the matching cost and aggregation weight to compensate the mismatch of focus. Experiments using both synthesized and real datasets demonstrate the effectiveness of our proposed algorithm.

****

Global Structure-From-Motion by Similarity Averaging

Zhaopeng Cui, Ping Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 864-872

Global structure-from-motion (SfM) methods solve all cameras simultaneously from all available relative motions. It has better potential in both reconstruction accuracy and computation efficiency than incremental methods. However, global SfM is challenging, mainly because of two reasons. Firstly, translation averaging is difficult, since an essential matrix only tells the direction of relative translation. Secondly, it is also hard to filter out bad essential matrices due to feature matching failures. We propose to compute a sparse depth image at each camera to solve both problems. Depth images help to upgrade an essential matrix to a similarity transformation, which can determine the scale of relative translation. Thus, camera registration is formulated as a well-posed similarity averaging problem. Depth images also make the filtering of essential matrices simple and effective. In this way, translation averaging can be solved robustly in two convex L1 optimization problems, which reach the global optimum rapidly. We demonstrate this method in various examples including sequential data, Internet data, and ambiguous data with repetitive scene structures.

****

Massively Parallel Multiview Stereopsis by Surface Normal Diffusion

Silvano Galliani, Katrin Lasinger, Konrad Schindler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 873-881

We present a new, massively parallel method for high-quality multiview matching. Our work builds on the Patchmatch idea: starting from randomly generated 3D planes in scene space, the best-fitting planes are iteratively propagated and refined to obtain a 3D depth and normal field per view, such that a robust photo-consistency measure over all images is maximized. Our main novelties are on the one hand to formulate Patchmatch in scene space, which makes it possible to aggregate image similarity across multiple views and obtain more accurate depth maps. And on the other hand a modified, diffusion-like propagation scheme that can be massively parallelized and delivers dense multiview correspondence over ten 1.9-Megapixel images in 3 seconds, on a consumer-grade GPU. Our method uses a slanted support window and thus has no fronto-parallel bias; it is completely local and parallel, such that computation time scales linearly with image size, and inversely proportional to the number of parallel threads. Furthermore, it has low memory footprint (four values per pixel, independent of the depth range). It therefore scales exceptionally well and can handle multiple large images at high depth resolution. Experiments on the DTU and Middlebury multiview datasets as well as oblique aerial images show that our method achieves very competitive results with high accuracy and completeness, across a range of different scenarios.

****

Variational PatchMatch MultiView Reconstruction and Refinement

Philipp Heise, Brian Jensen, Sebastian Klose, Alois Knoll; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 882-890

In this work we propose a novel approach to the problem of multi-view stereo rec onstruction. Building upon the previously proposed PatchMatch stereo and PM-Huber algorithm we introduce an extension to the multi-view scenario that employs an iterative refinement scheme. Our proposed approach uses an extended and robustified volumetric truncated signed distance function representation, which is advantageous for the fusion of refined depth maps and also for raycasting the current reconstruction estimation together with estimated depth normals into arbitrary camera views. We formulate the combined multi-view stereo reconstruction and refinement as a variational optimization problem. The newly introduced plane based smoothing term in the energy formulation is guided by the current reconstruction confidence and the image contents. Further we propose an extension of the PatchMatch scheme with an additional KLT step to avoid unnecessary sampling iterations. Improper camera poses are corrected by a direct image aligment step that performs robust outlier compensation by means of a recently proposed kernel lifting framework. To speed up the optimization of the variational formulation an adapted scheme is used for faster convergence.

****

As-Rigid-As-Possible Volumetric Shape-From-Template

Shaifali Parashar, Daniel Pizarro, Adrien Bartoli, Toby Collins; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 891-899

The objective of Shape-from-Template (SfT) is to infer an object's shape from a single image and a 3D object tem- plate. Existing methods are called thin-shell SfT as they represent the object by its outer surface. This may be an open surface for thin objects such as a piece of paper or a closed surface for thicker objects such as a ball. We pro- pose volumetric SfT, which specifically handles objects of the latter kind. Volumetric SfT uses the object's full volume to express the deformation constraints and reconstructs the object's surface and interior deformation. This is a chal- lenging problem because for opaque objects, only a part of the outer surface is visible in the image. Inspired by mesh- editing techniques, we use an As-Rigid-As-Possible (ARAP) deformation model that softly imposes local rigidity. We formalise ARAP isometric SfT as a constrained variational optimisation problem which we solve using iterative opti- misation. We present strategies to find an initial solution based on thin-shell SfT and volume propagation. Experi- ments with synthetic and real data show that our method has a typical maximum relative error of 5% in reconstruct- ing the deformation of an entire object, including its back and interior for which no visual data is available.

****

General Dynamic Scene Reconstruction From Multiple View Video

Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, Adrian Hilton; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 900-908

This paper introduces a general approach to dynamic scene   reconstruction from multiple moving cameras without prior knowledge or limiting constraints on the scene structure, appearance, or illumination. Existing techniques or dynamic scene reconstruction from multiple wide-baseline camera views primarily focus on accurate reconstruction in controlled environments, where the cameras are fixed and calibrated and background is known. These approaches are not robust for general dynamic scenes captured with sparse moving cameras. Previous approaches for outdoor dynamic scene reconstruction assume prior knowledge of the static background appearance and structure. The primary contributions of this paper are twofold: an automatic method for initial coarse dynamic scene segmentation and reconstruction without prior knowledge of background appearance or structure; and a general robust approach for joint segmentation refinement and dense reconstruction of dynamic scenes from multiple wide-baseline static or moving cameras. Evaluation is performed on a variety of indoor and outdoor scenes with cluttered backgrounds and multiple dynamic non-rigid objects such as people. Comparison with state-of-the-art approaches demonstrates improved accuracy in both multiple view segmentation and dense reconstruction. The proposed approach also eliminates the requirement for prior knowledge of scene structure and appearance.

****

The Joint Image Handbook

Matthew Trager, Martial Hebert, Jean Ponce; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 909-917

Given multiple perspective photographs, point correspondences form the "joint image", effectively a replica of three dimensional space distributed across its two-dimensional projections. This set can be characterized by multilinear equations over image coordinates, such as epipolar and trifocal constraints. We revisit in this paper the geometric and algebraic properties of the joint image, and address fundamental questions such as how many and which multilinearities are necessary and/or sufficient to determine camera geometry and/or image correspondences. The new theoretical results in this paper answer these questions in a very general setting and, in turn, are intended to serve as a "handbook" reference about multilinearities for practitioners.

****

Direct, Dense, and Deformable: Template-Based Non-Rigid 3D Reconstruction From RGB Video

Rui Yu, Chris Russell, Neill D. F. Campbell, Lourdes Agapito; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 918-926

In this paper we tackle the problem of capturing the dense, detailed 3D geometry of generic, complex non-rigid meshes using a single RGB-only commodity video camera and a direct approach. While robust and even real-time solutions exist to this problem if the observed scene is static, for non-rigid dense shape capture current systems are typically restricted to the use of complex multi-camera rigs, take advantage of the additional depth channel available in RGB-D cameras, or deal with specific shapes such as faces or planar surfaces. In contrast, our method makes use of a single RGB video as input; it can capture the deformations of generic shapes; and the depth estimation is dense, per-pixel and direct. We first compute a dense 3D template of the shape of the object, using a short rigid sequence, and subsequently perform online reconstruction of the non-rigid mesh as it evolves over time. Our energy optimization approach minimizes a robust photometric cost that simultaneously estimates the temporal correspondences and 3D deformations with respect to the template mesh. In our experimental evaluation we show a range of qualitative results on novel datasets; we compare against an existing method that requires multi-frame optical flow; and perform a quantitative evaluation against other template-based approaches on a ground truth dataset.

****

Single Image Pop-Up From Discriminatively Learned Parts

Menglong Zhu, Xiaowei Zhou, Kostas Daniilidis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 927-935

We introduce a new approach for estimating a fine grained 3D shape and continuous pose of an object from a single image. Given a training set of view exemplars, we learn and select appearance-based discriminative parts which are mapped onto the 3D model through a facility location optimization. The training set of 3D models is summarized into a set of basis shapes from which we can generalize by linear combination. Given a test image, we detect hypotheses for each part. The main challenge is to select from these hypotheses and compute the 3D pose and shape coefficients at the same time. To achieve this, we optimize a function that considers simultaneously the appearance matching of the parts as well as the geometric reprojection error. We apply the alternating direction method of multipliers (ADMM) to minimize the resulting convex function. Our main and novel contribution is the simultaneous solution for part localization and detailed 3D geometry estimation by maximizing both appearance and geometric compatibility with convex relaxation.

****

Learning Informative Edge Maps for Indoor Scene Layout Prediction

Arun Mallya, Svetlana Lazebnik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 936-944

In this paper, we introduce new edge-based features for the task of recovering the 3D layout of an indoor scene from a single image. Indoor scenes have certain edges that are very informative about the spatial layout of the room, namely, th e edges formed by the pairwise intersections of room faces (two walls, wall and ceiling, wall and floor). In contrast with previous approaches that rely on area-based features like geometric context and orientation maps, our method attempts to directly detect these informative edges. We learn to predict 'informative edge' probability maps using two recent methods that exploit local and global context, respectively: structured edge detection forests, and a fully convolutional network for pixelwise labeling. We show that the fully convolutional network is quite successful at predicting the informative edges even when they lack contrast or are occluded, and that the accuracy can be further improved by training the network to jointly predict the edges and the geometric context. Using features derived from the 'informative edge' maps, we learn a maximum margin structured classifier that achieves state-of-the-art performance on layout prediction.

****

Multi-View Convolutional Neural Networks for 3D Shape Recognition

Hang Su, Subhransu Maji, Evangelos Kalogerakis, Erik Learned-Miller; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 945-953

A longstanding question in computer vision concerns the representation of 3D shapes for recognition: should 3D shapes be represented with descriptors operating on their native 3D formats, such as voxel grid or polygon mesh, or can they be effectively represented with view-based descriptors? We address this question in the context of learning to recognize 3D shapes from a collection of their rendered views on 2D images. We first present a standard CNN architecture trained to recognize the shapes' rendered views independently of each other, and show that a 3D shape can be recognized even from a single view at an accuracy far higher than using state-of-the-art 3D shape descriptors. Recognition rates further increase when multiple views of the shapes are provided. In addition, we present a novel CNN architecture that combines information from multiple views of a 3D shape into a single and compact shape descriptor offering even better recognition performance. The same architecture can be applied to accurately recognize human hand-drawn sketches of shapes. We conclude that a collection of 2D views can be highly informative for 3D shape recognition and is amenable to emerging CNN architectures and their derivatives.

****

Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images

Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, Carsten Rother; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 954-962

Analysis-by-synthesis has been a successful approach for many tasks in computer vision, such as 6D pose estimation of an object in an RGB-D image which is the topic of this work. The idea is to compare the observation with the output of a forward process, such as a rendered image of the object of interest in a particular pose. Due to occlusion or complicated sensor noise, it can be difficult to perform this comparison in a meaningful way. We propose an approach that ``learns to compare'', while taking these difficulties into account. This is done by describing the posterior density of a particular object pose with a convolutional neural network (CNN) that compares observed and rendered images. The network is trained with the maximum likelihood paradigm. We observe empirically that the CNN does not specialize to the geometry or appearance of specific objects. It can be used with objects of vastly different shapes and appearances, and in different backgrounds. Compared to state-of-the-art, we demonstrate a significant improvement on two different datasets which include a total of eleven objects, cluttered background, and heavy occlusion.

****

3D Surface Profilometry Using Phase Shifting of De Bruijn Pattern

Matea Donlic, Tomislav Petkovic, Tomislav Pribanic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 963-971

A novel structured light method for color 3D surface profilometry is proposed. The proposed method does not require color calibration of a camera-projector pair and may be used for reconstruction of both dynamic and static scenes. The metho d uses a structured light pattern that is a combination of a De Bruijn color sequence and of a sinusoidal fringe. For dynamic scenes a Hessian ridge detector and a Gaussian mixture model are combined to extract stripe centers and to identify color. Stripes are then uniquely identified using dynamic programming based on the Smith-Waterman algorithm and a De Bruijn window property. For static scenes phase-shifting and De Bruijn window property are combined to obtain a high accuracy reconstruction. We have tested the proposed method on multiple objects with challenging surfaces and different albedos that demonstrate usability and robustness of the method.

****

A Deep Visual Correspondence Embedding Model for Stereo Matching Costs
Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, Chang Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 972-980
This paper presents a data-driven matching cost for stereo matching. A novel deep visual correspondence embedding model is trained via Convolutional Neural Network on a large set of stereo images with ground truth disparities. This deep embedding model leverages appearance data to learn visual similarity relationships between corresponding image patches, and explicitly maps intensity values into an embedding feature space to measure pixel dissimilarities. Experimental results on KITTI and Middlebury data sets demonstrate the effectiveness of our model. First, we prove that the new measure of pixel dissimilarity outperforms traditional matching costs. Furthermore, when integrated with a global stereo framework, our method ranks top 3 among all two-frame algorithms on the KITTI benchmark. Finally, cross-validation results show that our model is able to make correct predictions for unseen data which are outside of its labeled training set.

****

Learning Concept Embeddings With Combined Human-Machine Expertise
Michael Wilber, Iljung S. Kwak, David Kriegman, Serge Belongie; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 981-989
This paper presents our work on "SNaCK," a low-dimensional concept embedding algorithm that combines human expertise with automatic machine similarity kernels. Both parts are complimentary: human insight can capture relationships that are not apparent from the object's visual similarity and the machine can help relieve the human from having to exhaustively specify many constraints. We show that our SNaCK embeddings are useful in several tasks: distinguishing prime and nonprime numbers on MNIST, discovering labeling mistakes in the Caltech UCSD Birds (CUB) dataset with the help of deep-learned features, creating training datasets for bird classifiers, capturing subjective human taste on a new dataset of 10,000 foods, and qualitatively exploring an unstructured set of pictographic characters. Comparisons with the state-of-the-art in these tasks show that SNaCK produces better concept embeddings that require less human supervision than the leading methods.

****

Deep Multi-Patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation
Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, James Z. Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 990-998
This paper investigates problems of image style, aesthetics, and quality estimation, which require fine-grained details from high-resolution images, utilizing deep neural network training approach. Existing deep convolutional neural networks mostly extracted one patch such as a down-sized crop from each image as a training example. However, one patch may not always well represent the entire image, which may cause ambiguity during training. We propose a deep multi-patch aggregation network training approach, which allows us to train models using multiple patches generated from one image. We achieve this by constructing multiple, shared columns in the neural network and feeding multiple patches to each of the columns. More importantly, we propose two novel network layers (statistics and sorting) to support aggregation of those patches. The proposed deep multi-patch aggregation network integrates shared feature learning and aggregation function learning into a unified framework. We demonstrate the effectiveness of the deep mult

i-patch aggregation network on the three problems, i.e., image style recognition, aesthetic quality categorization, and image quality estimation. Our models trained using the proposed networks significantly outperformed the state of the art in all three applications.

****

Towards Computational Baby Learning: A Weakly-Supervised Approach for Object Detection
Xiaodan Liang, Si Liu, Yunchao Wei, Luoqi Liu, Liang Lin, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 999-1007
Intuitive observations show that a baby may inherently possess the capability of recognizing a new visual concept (e.g., chair, dog) by learning from only very few positive instances taught by parent(s) or others, and this recognition capability can be gradually further improved by exploring and/or interacting with the real instances in the physical world. Inspired by these observations, we propose a computational model for weakly-supervised object detection, based on prior knowledge modelling, exemplar learning and learning with video contexts. The prior knowledge is modeled with a pre-trained Convolutional Neural Network (CNN). When very few instances of a new concept are given, an initial concept detector is built by exemplar learning over the deep features the pre-trained CNN. The well-designed tracking solution is then used to discover more diverse instances from the massive online weakly labeled videos. Once a positive instance is detected/identified with high score in each video, more instances possibly from different view-angles and/or different distances are tracked and accumulated. Then the concept detector can be fine-tuned based on these new instances. This process can be repeated again and again till we obtain a very mature concept detector. Extensive experiments on Pascal VOC-07/10/12 object detection datasets well demonstrate the effectiveness of our framework. It can beat the state-of-the-art full-training based performances by learning from very few samples for each object category, along with about 20,000 weakly labeled videos.

****

Improving Image Classification With Location Context
Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, Lubomir Bourdev; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1008-1016
With the widespread availability of cellphones and cameras that have GPS capabilities, it is common for images being uploaded to the Internet today to have GPS coordinates associated with them. In addition to research that tries to predict GPS coordinates from visual features, this also opens up the door to problems that are conditioned on the availability of GPS coordinates. In this work, we tackle the problem of performing image classification with location context, in which we are given the GPS coordinates for images in both the train and test phases. We explore different ways of encoding and extracting features from the GPS coordinates, and show how to naturally incorporate these features into a Convolutional Neural Network (CNN), the current state-of-the-art for most image classification and recognition problems. We also show how it is possible to simultaneously learn the optimal pooling radii for a subset of our features within the CNN framework. To evaluate our model and to help promote research in this area, we identify a set of location-sensitive concepts and annotate a subset of the Yahoo Flickr Creative Commons 100M dataset that has GPS coordinates with these concepts, which we make publicly available. By leveraging location context, we are able to achieve almost a 7% gain in mean average precision.

****

HICO: A Benchmark for Recognizing Human-Object Interactions in Images
Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, Jia Deng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1017-1025
We introduce a new benchmark "Humans Interacting with Common Objects" (HICO) for recognizing human-object interactions (HOI). We demonstrate the key features of HICO: a diverse set of interactions with common object categories, a list of well-defined, sense-based HOI categories, and an exhaustive labeling of co-occurri

ng interactions with an object category in each image. We perform an in-depth analysis of representative current approaches and show that DNNs enjoy a significant edge. In addition, we show that semantic knowledge can significantly improve HOI recognition, especially for uncommon categories.

****

Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026-1034

Rectified activation units (rectifiers) are essential for state-of-the-art neural networks. In this work, we study rectifier neural networks for image classification from two aspects. First, we propose a Parametric Rectified Linear Unit (PReLU) that generalizes the traditional rectified unit. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk. Second, we derive a robust initialization method that particularly considers the rectifier nonlinearities. This method enables us to train extremely deep rectified models directly from scratch and to investigate deeper or wider network architectures. Based on the learnable activation and advanced initialization, we achieve 4.94% top-5 test error on the ImageNet 2012 classification dataset. This is a 26% relative improvement over the ILSVRC 2014 winner (GoogLeNet, 6.66%). To our knowledge, our result is the first to surpass the reported human-level performance (5.1%) on this dataset.

****

Continuous Pose Estimation With a Spatial Ensemble of Fisher Regressors

Michele Fenzi, Laura Leal-Taixe, Jorn Ostermann, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1035-1043

In this paper, we treat the problem of continuous pose estimation for object categories as a regression problem on the basis of only 2D training information. While regression is a natural framework for continuous problems, regression methods so far achieved inferior results with respect to 3D-based and 2D-based classification-and-refinement approaches. This may be attributed to their weakness to high intra-class variability as well as to noisy matching procedures and lack of geometrical constraints. We propose to apply regression to Fisher-encoded vectors computed from large cells by learning an array of Fisher regressors. Fisher encoding makes our algorithm flexible to variations in class appearance, while the array structure permits to indirectly introduce spatial context information in the approach. We formulate our problem as a MAP inference problem, where the likelihood function is composed of a generative term based on the prediction error generated by the ensemble of Fisher regressors as well as a discriminative term based on SVM classifiers. We test our algorithm on three publicly available datasets that envisage several difficulties, such as high intra-class variability, truncations, occlusions, and motion blur, obtaining state-of-the-art results.

****

Adaptive Hashing for Fast Similarity Search

Fatih Cakir, Stan Sclaroff; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1044-1052

With the staggering growth in image and video datasets, algorithms that provide fast similarity search and compact storage are crucial. Hashing methods that map the data into Hamming space have shown promise; however, many of these methods employ a batch-learning strategy in which the computational cost and memory requirements may become intractable and infeasible with larger and larger datasets. To overcome these challenges, we propose an online learning algorithm based on stochastic gradient descent in which the hash functions are updated iteratively with streaming data. In experiments with three image retrieval benchmarks, our online algorithm attains retrieval accuracy that is comparable to competing state-of-the-art batch-learning solutions, while our formulation is orders of magnitude faster and being online it is adaptable to the variations of the data. Moreover, our formulation yields improved retrieval performance over a recently reported online hashing technique, Online Kernel Hashing.

****

Single Image 3D Without a Single 3D Image

David F. Fouhey, Wajahat Hussain, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1053-1061

Do we really need 3D labels in order to learn how to predict 3D? In this paper, we show that one can learn a mapping from appearance to 3D properties without ever seeing a single explicit 3D label. Rather than use explicit supervision, we use the regularity of indoor scenes to learn the mapping in a completely unsupervised manner. We demonstrate this on both a standard 3D scene understanding dataset as well as Internet images for which 3D is unavailable, precluding supervised learning. Despite never seeing a 3D label, our method produces competitive results.

****

Cross-Domain Image Retrieval With a Dual Attribute-Aware Ranking Network

Junshi Huang, Rogerio S. Feris, Qiang Chen, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1062-1070

We address the problem of cross-domain image retrieval, considering the following practical application: given a user photo depicting a clothing image, our goal is to retrieve the same or attribute-similar clothing items from online shopping stores. This is a challenging problem due to the large discrepancy between online shopping images, usually taken in ideal lighting/pose/background conditions, and user photos captured in uncontrolled conditions. To address this problem, we propose a Dual Attribute-aware Ranking Network (DARN) for retrieval feature learning. More specifically, DARN consists of two sub-networks, one for each domain, whose retrieval feature representations are driven by semantic attribute learning. We show that this attribute-guided learning is a key factor for retrieval accuracy improvement. In addition, to further align with the nature of the retrieval problem, we impose a triplet visual similarity constraint for learning to rank across the two subnetworks. Another contribution of our work is a large-scale dataset which makes the network learning feasible. We exploit customer review websites to crawl a large set of online shopping images and corresponding offline user photos with fine-grained clothing attributes, i.e., around 450,000 online shopping images and about 90,000 exact offline counterpart images of those online ones. All these images are collected from real-world consumer websites reflecting the diversity of the data modality, which makes this dataset unique and rare in the academic community. We extensively evaluate the retrieval performance of networks in different configurations. The top-20 retrieval accuracy is doubled when using the proposed DARN other than the current popular solution using pre-trained CNN features only (0.570 vs. 0.268).

****

Attribute-Graph: A Graph Based Approach to Image Ranking

Nikita Prabhu, R. Venkatesh Babu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1071-1079

We propose a novel image representation, termed Attribute-Graph, to rank images by their semantic similarity to a given query image. An Attribute-Graph is an undirected fully connected graph, incorporating both local and global image characteristics. The graph nodes characterise objects as well as the overall scene context using mid-level semantic attributes, while the edges capture the object topology. We demonstrate the effectiveness of Attribute-Graphs by applying them to the problem of image ranking. We benchmark the performance of our algorithm on the 'rPascal' and 'rImageNet' datasets, which we have created in order to evaluate the ranking performance on complex queries containing multiple objects. Our experimental evaluation shows that modelling images as Attribute-Graphs results in improved ranking performance over existing techniques.

****

Contextual Action Recognition With R*CNN

Georgia Gkioxari, Ross Girshick, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1080-1088

There are multiple cues in an image which reveal what action a person is performing. For example, a jogger has a pose that is characteristic for jogging, but th e scene (e.g. road, trail) and the presence of other joggers can be an additional source of information. In this work, we exploit the simple observation that actions are accompanied by contextual cues to build a strong action recognition system. We adapt RCNN to use more than one region for classification while still maintaining the ability to localize the action. We call our system R*CNN. The action-specific models and the feature maps are trained jointly, allowing for action specific representations to emerge. R*CNN achieves 90.2% mean AP on the PASAL VOC Action dataset, outperforming all other approaches in the field by a significant margin. Last, we show that R*CNN is not limited to action recognition. In particular, R*CNN can also be used to tackle fine-grained tasks such as attribute classification. We validate this claim by reporting state-of-the-art performance on the Berkeley Attributes of People dataset.

****

What Makes an Object Memorable?

Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, Bernard Ghanem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1089-1097

Recent studies on image memorability have shed light on what distinguishes the memorability of different images and the intrinsic and extrinsic properties that make those images memorable. However, a clear understanding of the memorability of specific objects inside an image remains elusive. In this paper, we provide the first attempt to answer the question: what exactly is remembered about an image? We augment both the images and object segmentations from the PASCAL-S dataset with ground truth memorability scores and shed light on the various factors and properties that make an object memorable (or forgettable) to humans. We analyze various visual factors that may influence object memorability (e.g. color, visual saliency, and object categories). We also study the correlation between object and image memorability and find that image memorability is greatly affected by the memorability of its most memorable object. Lastly, we explore the effectiveness of deep learning and other computational approaches in predicting object memorability in images. Our efforts offer a deeper understanding of memorability in general thereby opening up avenues for a wide variety of applications.

****

kNN Hashing With Factorized Neighborhood Representation

Kun Ding, Chunlei Huo, Bin Fan, Chunhong Pan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1098-1106

Hashing is very effective for many tasks in reducing the processing time and in compressing massive databases. Although lots of approaches have been developed to learn data-dependent hash functions in recent years, how to learn hash functions to yield good performance with acceptable computational and memory cost is still a challenging problem. Based on the observation that retrieval precision is highly related to the kNN classification accuracy, this paper proposes a novel kNN-based supervised hashing method, which learns hash functions by directly maximizing the kNN accuracy of the Hamming-embedded training data. To make it scalable well to large problem, we propose a factorized neighborhood representation to parsimoniously model the neighborhood relationships inherent in training data. Considering that real-world data are often linearly inseparable, we further kernelize this basic model to improve its performance. As a result, the proposed method is able to learn accurate hashing functions with tolerable computation and storage cost. Experiments on four benchmarks demonstrate that our method outperforms the state-of-the-arts.

****

Multi-View Complementary Hash Tables for Nearest Neighbor Search

Xianglong Liu, Lei Huang, Cheng Deng, Jiwen Lu, Bo Lang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1107-1115

Recent years have witnessed the success of hashing techniques in fast nearest neighbor search. In practice many applications (e.g., visual search, object detection, image matching, etc.) have enjoyed the benefits of complementary hash tables and information fusion over multiple views. However, most of prior research mainly focused on compact hash code cleaning, and rare work studies how to build m ultiple complementary hash tables, much less to adaptively integrate information stemming from multiple views. In this paper we first present a novel multi-view complementary hash table method that learns complementarity hash tables from the data with multiple views. For single multi-view table, using exemplar based feature fusion, we approximate the inherent data similarities with a low-rank matrix, and learn discriminative hash functions in an efficient way. To build complementary tables and meanwhile maintain scalable training and fast out-of-sample extension, an exemplar reweighting scheme is introduced to update the induced low-rank similarity in the sequential table construction framework, which indeed brings mutual benefits between tables by placing greater importance on exemplars shared by mis-separated neighbors. Extensive experiments on three large-scale image datasets demonstrate that the proposed method significantly outperforms various naive solutions and state-of-the-art multi-table methods.

****

Scalable Person Re-Identification: A Benchmark

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116-1124

This paper contributes a new high quality dataset for person re-identification, named "Market-1501". Generally, current datasets: 1) are limited in scale; 2) consist of hand-drawn bboxes, which are unavailable under realistic settings; 3) have only one ground truth and one query image for each identity (close environment). To tackle these problems, the proposed Market-1501 dataset is featured in three aspects. First, it contains over 32,000 annotated bboxes, plus a distractor set of over 500K images, making it the largest person re-id dataset to date. Second, images in Market-1501 dataset are produced using the Deformable Part Model (DPM) as pedestrian detector. Third, our dataset is collected in an open system, where each identity has multiple images under each camera. As a minor contribution, inspired by recent advances in large-scale image search, this paper proposes an unsupervised Bag-of-Words descriptor. We view person re-identification as a special task of image search. In experiment, we show that the proposed descriptor yields competitive accuracy on VIPeR, CUHK03, and Market-1501 datasets, and is scalable on the large-scale 500k dataset.

****

MMSS: Multi-Modal Sharable and Specific Feature Learning for RGB-D Object Recognition

Anran Wang, Jianfei Cai, Jiwen Lu, Tat-Jen Cham; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1125-1133

Most of the feature-learning methods for RGB-D object recognition either learn features from color and depth modalities separately, or simply treat RGB-D as undifferentiated four-channel data, which cannot adequately exploit the relationship between different modalities. Motivated by the intuition that different modalities should contain not only some modal-specific patterns but also some shared common patterns, we propose a multi-modal feature learning framework for RGB-D object recognition. We first construct deep CNN layers for color and depth separately, and then connect them with our carefully designed multi-modal layers, which fuse color and depth information by enforcing a common part to be shared by features of different modalities. In this way, we obtain features reflecting shared properties as well as modal-specific properties in different modalities. The information of the multi-modal learning frameworks is back-propagated to the early CNN layers. Experimental results show that our proposed multi-modal feature learning method outperforms state-of-the-art approaches on two widely used RGB-D object benchmark datasets.

****

Object Detection via a Multi-Region and Semantic Segmentation-Aware CNN Model

Spyros Gidaris, Nikos Komodakis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1134-1142

We propose an object detection system that relies on a multi-region deep convolutional neural network (CNN) that also encodes semantic segmentation-aware features. The resulting CNN-based representation aims at capturing a diverse set of di scriminative appearance factors and exhibits localization sensitivity that is essential for accurate object localization. We exploit the above properties of our recognition module by integrating it on an iterative localization mechanism that alternates between scoring a box proposal and refining its location with a deep CNN regression model. Thanks to the efficient use of our modules, we detect objects with very high localization accuracy. On the detection challenges of PASCAL VOC2007 and PASCAL VOC2012 we achieve mAP of 78.2% and 73.9% correspondingly, surpassing any other published work by a significant margin.

****

Neural Activation Constellations: Unsupervised Part Model Discovery With Convolutional Networks

Marcel Simon, Erik Rodner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1143-1151

Part models of object categories are essential for challenging recognition tasks, where differences in categories are subtle and only reflected in appearances of small parts of the object. We present an approach that is able to learn part models in a completely unsupervised manner, without part annotations and even without given bounding boxes during learning. The key idea is to find constellations of neural activation patterns computed using convolutional neural networks. In our experiments, we outperform existing approaches for fine-grained recognition on the CUB200-2011, Oxford PETS, and Oxford Flowers dataset in case no part or bounding box annotations are available and achieve state-of-the-art performance for the Stanford Dog dataset. We also show the benefits of neural constellation models as a data augmentation technique for fine-tuning. Furthermore, our paper unites the areas of generic and fine-grained classification, since our approach is suitable for both scenarios.

****

Cascaded Sparse Spatial Bins for Efficient and Effective Generic Object Detection

David Novotny, Jiri Matas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1152-1160

A novel efficient method for extraction of object proposals is introduced. Its "objectness" function exploits deep spatial pyramid features, a novel fast-to-compute HoG-based edge statistic and the EdgeBoxes score. The efficiency is achieved by the use of spatial bins in a novel combination with sparsity-inducing group normalized SVM. State-of-the-art recall performance is achieved on Pascal VOC07, significantly outperforming methods with comparable speed. Interestingly, when only 100 proposals per image are considered the method attains 78 % recall on VOC07. The method improves mAP of the RCNN class-specific detector, increasing it by 10 points when only 50 proposals are used in each image. The system trained on twenty classes performs well on the two hundred class ILSVRC2013 set confirming generalization capability.

****

Probabilistic Label Relation Graphs With Ising Models

Nan Ding, Jia Deng, Kevin P. Murphy, Hartmut Neven; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1161-1169

We consider classification problems in which the label space has structure. A common example is hierarchical label spaces, corresponding to the case where one label subsumes another (e.g., animal subsumes dog). But labels can also be mutually exclusive (e.g., dog vs cat) or unrelated (e.g., furry, carnivore). To jointly model hierarchy and exclusion relations, the notion of a HEX (hierarchy and exclusion) graph was introduced. This combined a conditional random field (CRF) with a deep neural network (DNN), resulting in state of the art results when applied to visual object classification problems where the training labels were drawn from different levels of the ImageNet hierarchy (e.g., an image might be labeled with the basic level category "dog", rather than the more specific label "husky"). In this paper, we extend the HEX model to allow for soft or probabilistic relations between labels, which is useful when there is uncertainty about the relationship between two labels (e.g., an antelope is "sort of" furry, but not to the same degree as a grizzly bear). We call our new model pHEX, for probabilistic

HEX. We show that the pHEX graph can be converted to an Ising model, which allows us to use existing off-the-shelf inference methods (in contrast to the HEX method, which needed specialized inference algorithms). Experimental results show significant improvements in a number of large-scale visual object classification tasks, outperforming the previous HEX model.

****

Predicting Good Features for Image Geo-Localization Using Per-Bundle VLAD
Hyo Jin Kim, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1170-1178

We address the problem of recognizing a place depicted in a query image by using a large database of geo-tagged images at a city-scale. In particular, we discover features that are useful for recognizing a place in a data-driven manner, and use this knowledge to predict useful features in a query image prior to the geo-localization process. This allows us to achieve better performance while reducing the number of features. Also, for both learning to predict features and retrieving geo-tagged images from the database, we propose per-bundle vector of locally aggregated descriptors (PBVLAD), where each maximally stable region is described by a vector of locally aggregated descriptors (VLAD) on multiple scale-invariant features detected within the region. Experimental results show the proposed approach achieves a significant improvement over other baseline methods.

****

Task-Driven Feature Pooling for Image Classification
Guo-Sen Xie, Xu-Yao Zhang, Xiangbo Shu, Shuicheng Yan, Cheng-Lin Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1179-1187

Feature pooling is an important strategy to achieve high performance in image classification. However, most pooling methods are unsupervised and heuristic. In this paper, we propose a novel task-driven pooling (TDP) model to directly learn the pooled representation from data in a discriminative manner. Different from the traditional methods (e.g., average and max pooling), TDP is an implicit pooling method which elegantly integrates the learning of representations into the given classification task. The optimization of TDP can equalize the similarities between the descriptors and the learned representation, and maximize the classification accuracy. TDP can be combined with the traditional BoW models (coding vectors) or the recent state-of-the-art CNN models (feature maps) to achieve a much better pooled representation. Furthermore, a self-training mechanism is used to generate the TDP representation for a new test image. A multi-task extension of TDP is also proposed to further improve the performance. Experiments on three databases (Flower-17, Indoor-67 and Caltech-101) well validate the effectiveness of our models.

****

Cutting Edge: Soft Correspondences in Multimodal Scene Parsing
Sarah Taghavi Namin, Mohammad Najafi, Mathieu Salzmann, Lars Petersson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1188-1196

Exploiting multiple modalities for semantic scene parsing has been shown to improve accuracy over the single modality scenario. Existing methods, however, assume that corresponding regions in two modalities have the same label. In this paper, we address the problem of data misalignment and label inconsistencies, e.g., due to moving objects, in semantic labeling, which violate the assumption of existing techniques. To this end, we formulate multimodal semantic labeling as inference in a CRF, and introduce latent nodes to explicitly model inconsistencies between two domains. These latent nodes allow us not only to leverage information from both domains to improve their labeling, but also to cut the edges between inconsistent regions. To eliminate the need for hand tuning the parameters of our model, we propose to learn intra-domain and inter-domain potential functions from training data. We demonstrate the benefits of our approach on two publicly available datasets containing 2D imagery and 3D point clouds. Thanks to our latent nodes and our learning strategy, our method outperforms the state-of-the-art in both cases.

****

One Shot Learning via Compositions of Meaningful Patches

Alex Wong, Alan L. Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1197-1205

The task of discriminating one object from another is almost trivial for a human being. However, this task is computationally taxing for most modern machine learning methods; whereas, we perform this task at ease given very few examples for learning. It has been proposed that the quick grasp of concept may come from the shared knowledge between the new example and examples previously learned. We believe that the key to one-shot learning is the sharing of common parts as each part holds immense amounts of information on how a visual concept is constructed. We propose an unsupervised method for learning a compact dictionary of image patches representing meaningful components of an objects. Using those patches as features, we build a compositional model that outperforms a number of popular algorithms on a one-shot learning task. We demonstrate the effectiveness of this approach on hand-written digits and show that this model generalizes to multiple datasets.

****

FASText: Efficient Unconstrained Scene Text Detector

Michal Busta, Lukas Neumann, Jiri Matas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1206-1214

We propose a novel easy-to-implement stroke detector based on an efficient pixel intensity comparison to surrounding pixels. Stroke-specific keypoints are efficiently detected and text fragments are subsequently extracted by local thresholding guided by keypoint properties. Classification based on effectively calculated features then eliminates non-text regions. The stroke-specific keypoints produce 2 times less region segmentations and still detect 25% more characters than the commonly exploited MSER detector and the process is 4 times faster. After a novel efficient classification step, the number of regions is reduced to 7 times less than the standard method and is still almost 3 times faster. All stages of the proposed pipeline are scale- and rotation-invariant and support a wide variety of scripts (Latin, Hebrew, Chinese, etc.) and fonts. When the proposed detector is plugged into a scene text localization and recognition pipeline, a state-of-the-art text localization accuracy is maintained whilst the processing time is significantly reduced.

****

Multi-Scale Recognition With DAG-CNNs

Songfan Yang, Deva Ramanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1215-1223

We explore multi-scale convolutional neural nets (CNNs) for image classification. Contemporary approaches extract features from a single output layer. By extracting features from multiple layers, one can simultaneously reason about high, mid, and low-level features during classification. The resulting multi-scale architecture can itself be seen as a feed-forward model that is structured as a directed acyclic graph (DAG-CNNs). We use DAG-CNNs to learn a set of multi-scale features that can be effectively shared between coarse and fine-grained classification tasks. While fine-tuning such models helps performance, we show that even "off-the-self" multi-scale features perform quite well. We present extensive analysis and demonstrate state-of-the-art classification performance on three standard scene benchmarks (SUN397, MIT67, and Scene15). In terms of the heavily benchmarked MIT67 and Scene15 datasets, our results reduce the lowest previously-reported error by 23.9% and 9.5%, respectively.

****

Relaxed Multiple-Instance SVM With Application to Object Discovery

Xinggang Wang, Zhuotun Zhu, Cong Yao, Xiang Bai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1224-1232

Multiple-instance learning (MIL) has served as an important tool for a wide range of vision applications, for instance, image classification, object detection, and visual tracking. In this paper, we propose a novel method to solve the classical MIL problem, named relaxed multiple-instance SVM (RMI-SVM). We treat the po sitiveness of instance as a continuous variable, use Noisy-OR model to enforce the MIL constraints, and optimize them jointly in a unified framework. The optimization problem can be efficiently solved using stochastic gradient decent. The extensive experiments demonstrate that RMI-SVM consistently achieves superior performance on various benchmarks for MIL. Moreover, we simply applied RMI-SVM to a challenging vision task, common object discovery. The state-of-the arts results of object discovery on PASCAL VOC datasets further confirm the advantages of the proposed method.

****

Im2Calories: Towards an Automated Mobile Vision Food Diary
Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, Kevin P. Murphy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1233-1241

We present a system which can recognize the contents of your meal from a single image, and then predict its nutritional contents, such as calories. The simplest version assumes that the user is eating at a restaurant for which we know the menu. In this case, we can collect images offline to train a multi-label classifier. At run time, we apply the classifier (running on your phone) to predict which foods are present in your meal, and we lookup the corresponding nutritional facts. We apply this method to a new dataset of images from 23 different restaurants, using a CNN-based classifier, significantly outperforming previous work. The more challenging setting works outside of restaurants. In this case, we need to estimate the size of the foods, as well as their labels. This requires solving segmentation and depth / volume estimation from a single image. We present CNN-based approaches to these problems, with promising preliminary results.

****

LEWIS: Latent Embeddings for Word Images and their Semantics
Albert Gordo, Jon Almazan, Naila Murray, Florent Perronin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1242-1250

The goal of this work is to bring semantics into the tasks of text recognition and retrieval in natural images. Although text recognition and retrieval have received a lot of attention in recent years, previous works have focused on recognizing or retrieving exactly the same word used as a query, without taking the
In this paper, we ask the following question: can we predict semantic concepts directly from a word image, without explicitly trying to transcribe the word image or its characters at any point? For this goal we propose a convolutional neural network (CNN) with a weighted ranking loss objective that ensures that the concepts relevant to the query image are ranked ahead of those that are not relevant. This can also be interpreted as learning a Euclidean space where word images and concepts are jointly embedded. This model is learned in an end-to-end manner, from image pixels to semantic concepts, using a dataset of synthetically generated word images and concepts mined from a lexical database (WordNet). Our results show that, despite the complexity of the task, word images and concepts can indeed be associated with a high degree of accuracy.

****

Per-Sample Kernel Adaptation for Visual Recognition and Grouping
Borislav Antic, Bjorn Ommer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1251-1259

Object, action, or scene representations that are corrupted by noise significantly impair the performance of visual recognition. Typically, partial occlusion, clutter, or excessive articulation affects only a subset of all feature dimensions and, most importantly, different dimensions are corrupted in different samples. Nevertheless, the common approach to this problem in feature selection and kernel methods is to down-weight or eliminate entire training samples or the same dimensions of all samples. Thus, valuable signal is lost, resulting in suboptimal classification. Our goal is, therefore, to adjust the contribution of individual feature dimensions when comparing any two samples and computing their similarity. Consequently, per-sample selection of informative dimensions is directly in tegrated into kernel computation. The interrelated problems of learning the parameters of a kernel classifier and determining the informative components of each sample are then addressed in a joint objective function. The approach can be integrated into the learning stage of any kernel-based visual recognition problem and it does not affect the computational performance in the retrieval phase. Experiments on diverse challenges of action recognition in videos and indoor scene classification show the general applicability of the approach and its ability to improve learning of visual representations.

****

Fine-Grained Change Detection of Misaligned Scenes With Varied Illuminations

Wei Feng, Fei-Peng Tian, Qian Zhang, Nan Zhang, Liang Wan, Jizhou Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1260-1268

Detecting fine-grained subtle changes among a scene is critically important in practice. Previous change detection methods, focusing on detecting large-scale significant changes, cannot do this well. This paper proposes a feasible end-to-end approach to this challenging problem. We start from active camera relocation that quickly relocates camera to nearly the same pose and position of the last time observation. To guarantee detection sensitivity and accuracy of minute changes, in an observation, we capture a group of images under multiple illuminations, which need only to be roughly aligned to the last time lighting conditions. Given two times observations, we formulate fine-grained change detection as a joint optimization problem of three related factors, i.e., normal-aware lighting difference, camera geometry correction flow, and real scene change mask. We solve the three factors in a coarse-to-fine manner and achieve reliable change decision by rank minimization. We build three real-world datasets to benchmark fine-grained change detection of misaligned scenes under varied multiple lighting conditions. Extensive experiments show the superior performance of our approach over state-of-the-art change detection methods and its ability to distinguish real scene changes from false ones caused by lighting variations.

****

Aggregating Local Deep Features for Image Retrieval

Artem Babenko, Victor Lempitsky; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1269-1277

Several recent works have shown that image descriptors produced by deep convolutional neural networks provide state-of-the-art performance for image classification and retrieval problems. It also has been shown that the activations from the convolutional layers can be interpreted as local features describing particular image regions. These local features can be aggregated using aggregating methods developed for local features (e.g. Fisher vectors), thus providing new powerful global descriptor. In this paper we investigate possible ways to aggregate local deep features to produce compact descriptors for image retrieval. First, we show that deep features and traditional hand-engineered features have quite different distributions of pairwise similarities, hence existing aggregation methods have to be carefully re-evaluated. Such re-evaluation reveals that in contrast to shallow features, the simple aggregation method based on sum pooling provides the best performance for deep convolutional features. This method is efficient, has few parameters, and bears little risk of overfitting when e.g. learning the PCA matrix. In addition, we suggest a simple yet efficient query expansion scheme suitable for the proposed aggregation method. Overall, the new compact global descriptor improves the state-of-the-art on four common benchmarks considerably.

****

Learning Deep Object Detectors From 3D Models

Xingchao Peng, Baochen Sun, Karim Ali, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1278-1286

Crowdsourced 3D CAD models are easily accessible online, and can potentially generate an infinite number of training images for almost any object category. We show that augmenting the training data of contemporary Deep Convolutional Neural Net (DCNN) models with such synthetic data can be effective, especially when real training data is limited or not well matched to the target domain. Most freely available CAD models capture 3D shape but are often missing other low level cues, such as realistic object texture, pose, or background. In a detailed analysis, we use synthetic CAD images to probe the ability of DCNN to learn without these cues, with surprising findings. In particular, we show that when the DCNN is fine-tuned on the target detection task, it exhibits a large degree of invariance to missing low-level cues, but, when pretrained on generic ImageNet classification, it learns better when the low-level cues are simulated. We show that our synthetic DCNN training approach significantly outperforms previous methods on the benchmark PASCAL VOC2007 dataset when learning in the few-shot scenario and improves performance in a domain shift scenario on the Office benchmark.

****

Harvesting Discriminative Meta Objects With Deep CNN Features for Scene Classification

Ruobing Wu, Baoyuan Wang, Wenping Wang, Yizhou Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1287-1295

Recent work on scene classification still makes use of generic CNN features in a rudimentary manner. In this paper, we present a novel pipeline built upon deep CNN features to harvest discriminative visual objects and parts for scene classification. We first use a region proposal technique to generate a set of high-quality patches potentially containing objects, and apply a pre-trained CNN to extract generic deep features from these patches. Then we perform both unsupervised and weakly supervised learning to screen these patches and discover discriminative ones representing category-specific objects and parts. We further apply discriminative clustering enhanced with local CNN fine-tuning to aggregate similar objects and parts into groups, called meta objects. A scene image representation is constructed by pooling the feature response maps of all the learned meta objects at multiple spatial scales. We have confirmed that the scene image representation obtained using this new pipeline is capable of delivering state-of-the-art performance on two popular scene benchmark datasets, MIT Indoor 67 [22] and Sun397 [31].

****

Scalable Nonlinear Embeddings for Semantic Category-Based Image Retrieval

Gaurav Sharma, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1296-1304

We propose a novel algorithm for the task of supervised discriminative distance learning by nonlinearly embedding vectors into a low dimensional Euclidean space. We work in the challenging setting where supervision is with constraints on similar and dissimilar pairs while training. The proposed method is derived by an approximate kernelization of a linear Mahalanobis-like distance metric learning algorithm and can also be seen as a kernel neural network. The number of model parameters and test time evaluation complexity of the proposed method are O(dD) where D is the dimensionality of the input features and d is the dimension of the projection space -- this is in contrast to the usual kernelization methods as, unlike them, the complexity does not scale linearly with the number of training examples. We propose a stochastic gradient based learning algorithm which makes the method scalable (w.r.t. the number of training examples), while being nonlinear. We train the method with up to half a million training pairs of 4096 dimensional CNN features. We give empirical comparisons with relevant baselines on seven challenging datasets for the task of low dimensional semantic category based image retrieval.

****

Person Re-Identification Ranking Optimisation by Discriminant Context Information Analysis

Jorge Garcia, Niki Martinel, Christian Micheloni, Alfredo Gardel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1305-1313

Person re-identification is an open and challenging problem in computer vision. Existing re-identification approaches focus on optimal methods for features matching (e.g., metric learning approaches) or study the inter-camera transformations of such features. These methods hardly ever pay attention to the problem of vi sual ambiguities shared between the first ranks. In this paper, we focus on such a problem and introduce an unsupervised ranking optimization approach based on discriminant context information analysis. The proposed approach refines a given initial ranking by removing the visual ambiguities common to first ranks. This is achieved by analyzing their content and context information. Extensive experiments on three publicly available benchmark datasets and different baseline methods have been conducted. Results demonstrate a remarkable improvement in the first positions of the ranking. Regardless of the selected dataset, state-of-the-art methods are strongly outperformed by our method.

****

Unsupervised Generation of a Viewpoint Annotated Car Dataset From Videos

Nima Sedaghat, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1314-1322

Object recognition approaches have recently been extended to yield, aside of the object class output, also viewpoint or pose. Training such approaches typically requires additional viewpoint or keypoint annotation in the training data or, alternatively, synthetic CAD models. In this paper,we present an approach that creates a dataset of images annotated with bounding boxes and viewpoint labels in a fully automated manner from videos. We assume that the scene is static in order to reconstruct 3D surfaces via structure from motion. We automatically detect when the reconstruction fails and normalize for the viewpoint of the 3D models by aligning the reconstructed point clouds. Exemplarily for cars we show that we can expand a large dataset of annotated single images and obtain improved performance when training a viewpoint regressor on this joined dataset.

****

Structured Indoor Modeling

Satoshi Ikehata, Hang Yang, Yasutaka Furukawa; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1323-1331

This paper presents a novel 3D modeling framework that reconstructs an indoor scene as a structured model from panorama RGBD images. A scene geometry is represented as a graph, where nodes correspond to structural elements such as rooms, walls, and objects. The approach devises a structure grammar that defines how a scene graph can be manipulated. The grammar then drives a principled new reconstruction algorithm, where the grammar rules are sequentially applied to recover a structured model. The paper also proposes a new room segmentation algorithm and an offset-map reconstruction algorithm that are used in the framework and can enforce architectural shape priors far beyond existing state-of-the-art. The structured scene representation enables a variety of novel applications, ranging from indoor scene visualization, automated floorplan generation, Inverse-CAD, and more. We have tested our framework and algorithms on six synthetic and five real datasets with qualitative and quantitative evaluations.

****

3D Time-Lapse Reconstruction From Internet Photos

Ricardo Martin-Brualla, David Gallup, Steven M. Seitz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1332-1340

Given an Internet photo collection of a landmark, we compute a 3D time-lapse video sequence where a virtual camera moves continuously in time and space. While previous work assumed a static camera, the addition of camera motion during the time-lapse creates a very compelling impression of parallax. Achieving this goal, however, requires addressing multiple technical challenges, including solving for time-varying depth maps, regularizing 3D point color profiles over time, and reconstructing high quality, hole-free images at every frame from the projected profiles. Our results show photorealistic time-lapses of skylines and natural scenes over many years, with dramatic parallax effects.

****

Global, Dense Multiscale Reconstruction for a Billion Points

Benjamin Ummenhofer, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1341-1349

We present a variational approach for surface reconstruction from a set of oriented points with scale information. We focus particularly on scenarios with non-u niform point densities due to images taken from different distances. In contrast to previous methods, we integrate the scale information in the objective and globally optimize the signed distance function of the surface on a balanced octree grid. We use a finite element discretization on the dual structure of the octree minimizing the number of variables. The tetrahedral mesh is generated efficiently from the dual structure, and also memory efficiency is optimized, such that robust data terms can be used even on very large scenes. The surface normals are explicitly optimized and used for surface extraction to improve the reconstruction at edges and corners.

****

On the Visibility of Point Clouds

Sagi Katz, Ayellet Tal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1350-1358

Is it possible to determine the visible subset of points directly from a given point cloud? Interestingly, in [7] it was shown that this is indeed the case - despite the fact that points cannot occlude each other, this task can be performed without surface reconstruction or normal estimation. The operator is very simple - it first transforms the points to a new domain and then constructs the convex hull in that domain. Points that lie on the convex hull of the transformed set of points are the images of the visible points. This operator found numerous applications in computer vision, including face reconstruction, keypoint detection, finding the best viewpoints, reduction of points, and many more. The current paper addresses a fundamental question: What properties should a transformation function satisfy, in order to be utilized in this operator? We show that three such properties are sufficient: the sign of the function, monotonicity, and a condition regarding the function's parameter. The correctness of an algorithm that satisfies these three properties is proved. Finally, we show an interesting application of the operator - assignment of visibility-confidence score. This feature is missing from previous approaches, where a binary yes/no visibility is determined. This score can be utilized in various applications; we illustrate its use in view-dependent curvature estimation.

****

Weakly Supervised Graph Based Semantic Segmentation by Learning Communities of Image-Parts

Niloufar Pourian, S. Karthikeyan, B.S. Manjunath; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1359-1367

We present a weakly-supervised approach to semantic segmentation. The goal is to assign pixel-level labels given only partial information, for example, image-level labels. This is an important problem in many application scenarios where it is difficult to get accurate segmentation or not feasible to obtain detailed annotations. The proposed approach starts with an initial coarse segmentation, followed by a spectral clustering approach that groups related image parts into communities. A community-driven graph is then constructed that captures spatial and feature relationships between communities while a label graph captures correlations between image labels. Finally, mapping the image level labels to appropriate communities is formulated as a convex optimization problem. The proposed approach does not require location information for image level labels and can be trained using partially labeled datasets. Compared to the state-of-the-art weakly supervised approaches, we achieve a significant performance improvement of 9% on MSRC-21 dataset and 11% on LabelMe dataset, while being more than 300 times faster.

****

Piecewise Flat Embedding for Image Segmentation

Yizhou Yu, Chaowei Fang, Zicheng Liao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1368-1376

Image segmentation is a critical step in many computer vision tasks, including high-level visual recognition and scene understanding as well as low-level photo and video processing. In this paper, we propose a new nonlinear embedding, called piecewise flat embedding, for image segmentation. Based on the theory of sparse signal recovery, piecewise flat embedding attempts to identify segment boundar ies while significantly suppressing variations within segments. We adopt an L1-regularized energy term in the formulation to promote sparse solutions. We further devise an effective two-stage numerical algorithm based on Bregman iterations to solve the proposed embedding. Piecewise flat embedding can be easily integrated into existing image segmentation frameworks, including segmentation based on spectral clustering and hierarchical segmentation based on contour detection. Experiments on BSDS500 indicate that segmentation algorithms incorporating this embedding can achieve significantly improved results in both frameworks.

****

Semantic Image Segmentation via Deep Parsing Network

Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1377-1385

This paper addresses semantic image segmentation by incorporating rich information into Markov Random Field (MRF), including high-order relations and mixture of label contexts. Unlike previous works that optimized MRFs using iterative algorithm, we solve MRF by proposing a Convolutional Neural Network (CNN), namely Deep Parsing Network (DPN), which enables deterministic end-to-end computation in a single forward pass. Specifically, DPN extends a contemporary CNN architecture to model unary terms and additional layers are carefully devised to approximate the mean field algorithm (MF) for pairwise terms. It has several appealing properties. First, different from the recent works that combined CNN and MRF, where many iterations of MF were required for each training image during back-propagation, DPN is able to achieve high performance by approximating one iteration of MF. Second, DPN represents various types of pairwise terms, making many existing works as its special cases. Third, DPN makes MF easier to be parallelized and speeded up in Graphical Processing Unit (GPU). DPN is thoroughly evaluated on the PASCAL VOC 2012 dataset, where a single DPN model yields a new state-of-the-art segmentation accuracy of 77.5%.

****

Human Parsing With Contextualized Convolutional Neural Network

Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1386-1394

In this work, we address the human parsing task with a novel Contextualized Convolutional Neural Network (Co-CNN) architecture, which well integrates the cross-layer context, global image-level context, within-super-pixel context and cross-super-pixel neighborhood context into a unified network. Given an input human image, Co-CNN produces the pixel-wise categorization in an end-to-end way. First, the cross-layer context is captured by our basic local-to-global-to-local structure, which hierarchically combines the global semantic structure and the local fine details within the cross-layers. Second, the global image-level label prediction is used as an auxiliary objective in the intermediate layer of the Co-CNN, and its outputs are further used for guiding the feature learning in subsequent convolutional layers to leverage the global image-level context. Finally, to further utilize the local super-pixel contexts, the within-super-pixel smoothing and cross-super-pixel neighbourhood voting are formulated as natural sub-components of the Co-CNN to achieve the local label consistency in both training and testing process. Comprehensive evaluations on two public datasets well demonstrate the significant superiority of our Co-CNN architecture over other state-of-the-arts for human parsing. In particular, the F-1 score on the large dataset reaches 76.95% by Co-CNN, significantly higher than 62.81% and 64.38% by the state-of-the-art algorithms, M-CNN and ATR, respectively.

****

Holistically-Nested Edge Detection

Saining Xie, Zhuowen Tu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1395-1403

We develop a new edge detection algorithm that addresses two critical issues in this long-standing vision problem: (1) holistic image training; and (2) multi-scale feature learning. Our proposed method, holistically-nested edge detection (HED), turns pixel-wise edge classification into image-to-image prediction by mean s of a deep learning model that leverages fully convolutional neural networks and deeply-supervised nets. HED automatically learns rich hierarchical representations (guided by deep supervision on side responses) that are crucially important in order to approach the human ability to resolve the challenging ambiguity in edge and object boundary detection. We significantly advance the state-of-the-art on the BSD500 dataset (ODS F-score of 0.782) and the NYU Depth dataset (ODS F-score of 0.746), and do so with an improved speed (0.4 second per image) that is orders of magnitude faster than recent CNN-based edge detection algorithms.

****

Minimum Barrier Salient Object Detection at 80 FPS

Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, Radomir Mech; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1404-1412

We propose a highly efficient, yet powerful, salient object detection method based on the Minimum Barrier Distance (MBD) Transform. The MBD transform is robust to pixel-value fluctuation, and thus can be effectively applied on raw pixels without region abstraction. We present an approximate MBD transform algorithm with 100X speedup over the exact algorithm. An error bound analysis is also provided. Powered by this fast MBD transform algorithm, the proposed salient object detection method runs at 80 FPS, and significantly outperforms previous methods with similar speed on four large benchmark datasets, and achieves comparable or better performance than state-of-the-art methods. Furthermore, a technique based on color whitening is proposed to extend our method to leverage the appearance-based backgroundness cue. This extended version further improves the performance, while still being one order of magnitude faster than all the other leading methods.

****

Learning Image Representations Tied to Ego-Motion

Dinesh Jayaraman, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1413-1421

Understanding how images of objects and scenes behave in response to specific ego-motions is a crucial aspect of proper visual development, yet existing visual learning methods are conspicuously disconnected from the physical source of their images. We propose to exploit proprioceptive motor signals to provide unsupervised regularization in convolutional neural networks to learn visual representations from egocentric video. Specifically, we enforce that our learned features exhibit equivariance, i.e, they respond predictably to transformations associated with distinct ego-motions. With three datasets, we show that our unsupervised feature learning approach significantly outperforms previous approaches on visual recognition and next-best-view prediction tasks. In the most challenging test, we show that features learned from video captured on an autonomous driving platform improve large-scale scene recognition in static images from a disjoint domain.

****

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei A. Efros; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1422-1430

This work explores the use of spatial context as a source of free and plentiful supervisory signal for training a rich visual representation. Given only a large, unlabeled image collection, we extract random pairs of patches from each image and train a convolutional neural net to predict the position of the second patch relative to the first. We argue that doing well on this task requires the model to learn to recognize objects and their parts. We demonstrate that the feature representation learned using this within-image context indeed captures visual similarity across images. For example, this representation allows us to perform unsupervised visual discovery of objects like cats, people, and even birds from the Pascal VOC 2011 detection dataset. Furthermore, we show that the learned ConvNet can be used in the R-CNN framework and provides a significant boost over a randomly-initialized ConvNet, resulting in state-of-the-art performance among algorithms which use only Pascal-provided training set annotations.

****

Webly Supervised Learning of Convolutional Networks

Xinlei Chen, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1431-1439

We present an approach to utilize large amounts of web data for learning CNNs. Specifically inspired by curriculum learning, we present a two-step approach for CNN training. First, we use easy images to train an initial visual representation. We then use this initial CNN and adapt it to harder, more realistic images by leveraging the structure of data and categories. We demonstrate that our two-stage CNN outperforms a fine-tuned CNN trained on ImageNet on Pascal VOC 2012. We also demonstrate the strength of webly supervised learning by localizing objects in web images and training a R-CNN style detector. It achieves the best performance on VOC 2007 where no VOC training data is used. Finally, we show our approach is quite robust to noise and performs comparably even when we use image search results from March 2013 (pre-CNN image search era).

****

Fast R-CNN

Ross Girshick; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448

This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. Fast R-CNN trains the very deep VGG16 network 9x faster than R-CNN, is 213x faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3x faster, tests 10x faster, and is more accurate. Fast R-CNN is implemented in Python and C++ (using Caffe) and is available under the open-source MIT License at https://github.com/rbgirshick/fast-rcnn.

****

Bilinear CNN Models for Fine-Grained Visual Recognition

Tsung-Yu Lin, Aruni RoyChowdhury, Subhransu Maji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1449-1457

We propose bilinear models, a recognition architecture that consists of two feature extractors whose outputs are multiplied using outer product at each location of the image and pooled to obtain an image descriptor. This architecture can model local pairwise feature interactions in a translationally invariant manner which is particularly useful for fine-grained categorization. It also generalizes various orderless texture descriptors such as the Fisher vector, VLAD and O2P. We present experiments with bilinear models where the feature extractors are based on convolutional neural networks. The bilinear form simplifies gradient computation and allows end-to-end training of both networks using image labels only. Using networks initialized from the ImageNet dataset followed by domain specific fine-tuning we obtain 84.1% accuracy of the CUB-200-2011 dataset requiring only category labels at training time. We present experiments and visualizations that analyze the effects of fine-tuning and the choice two networks on the speed and accuracy of the models. Results show that the architecture compares favorably to the existing state of the art on a number of fine-grained datasets while being substantially simpler and easier to train. Moreover, our most accurate model is fairly efficient running at 8 frames/sec on a NVIDIA Tesla K40 GPU. The source code for the complete system will be made available at http://vis-www.cs.umass.edu/bcnn

****

Discovering the Spatial Extent of Relative Attributes

Fanyi Xiao, Yong Jae Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1458-1466

We present a weakly-supervised approach that discovers the spatial extent of relative attributes, given only pairs of ordered images. In contrast to traditional approaches that use global appearance features or rely on keypoint detectors, our goal is to automatically discover the image regions that are relevant to the attribute, even when the attribute's appearance changes drastically across its attribute spectrum. To accomplish this, we first develop a novel formulation that combines a detector with local smoothness to discover a set of coherent visual chains across the image collection. We then introduce an efficient way to generate additional chains anchored on the initial discovered ones. Finally, we automatically identify the most relevant visual chains, and create an ensemble image representation to model the attribute. Through extensive experiments, we demonstrate our method's promise relative to several baselines in modeling relative attributes.

****

Deep Neural Decision Forests

Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, Samuel Rota Bulo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1467-1475

We present Deep Neural Decision Forests - a novel approach that unifies classification trees with the representation learning functionality known from deep convolutional networks, by training them in an end-to-end manner. To combine these two worlds, we introduce a stochastic and differentiable decision tree model, which steers the representation learning usually conducted in the initial layers of a (deep) convolutional network. Our model differs from conventional deep networks because a decision forest provides the final predictions and it differs from conventional decision forests since we propose a principled, joint and global optimization of split and leaf node parameters. We show experimental results on benchmark machine learning datasets like MNIST and ImageNet and find on-par or superior results when compared to state-of-the-art deep models. Most remarkably, we obtain Top5-Errors of only 7.84%/6.38% on ImageNet validation data when integrating our forests in a single-crop, single/seven model GoogLeNet architecture, respectively. Thus, even without any form of training data set augmentation we are improving on the 6.67% error obtained by the best GoogLeNet architecture (7 models, 144 crops).

****

Deep Fried Convnets

Zichao Yang, Marcin Moczulski, Misha Denil, Nando de Freitas, Alex Smola, Le Song, Ziyu Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1476-1483

The fully connected layers of a deep convolutional neural network typically contain over 90% of the network parameters, and consume the majority of the memory required to store the network. Reducing the number of parameters while preserving predictive performance is critically important for deploying deep neural networks in memory constrained environments such as GPUs or embedded devices. In this paper we show how kernel methods, in particular a single Fastfood layer, can be used to replace the fully connected layers in a deep convolutional neural network. This deep fried network is end-to-end trainable in conjunction with convolutional layers. Our new architecture substantially reduces the memory footprint of convolutional networks trained on MNIST and ImageNet with no drop in predictive performance

****

Semantic Component Analysis

Calvin Murdock, Fernando De la Torre; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1484-1492

Unsupervised and weakly-supervised visual learning in large image collections are critical in order to avoid the time-consuming and error-prone process of manual labeling. Standard approaches rely on methods like multiple-instance learning or graphical models, which can be computationally intensive and sensitive to initialization. On the other hand, simpler component analysis or clustering methods usually cannot achieve meaningful invariances or semantic interpretability. To address the issues of previous work, we present a simple but effective method called Semantic Component Analysis (SCA), which provides a decomposition of images into semantic components. Unsupervised SCA decomposes additive image representations into spatially-meaningful visual components that naturally correspond to object categories. Using an overcomplete representation that allows for rich instance-level constraints and spatial priors, SCA gives improved results and more interpretable components in comparison to traditional matrix factorization techniques. If weakly-supervised information is available in the form of image-level tags, SCA factorizes a set of images into semantic groups of superpixels. We also provide qualitative connections to traditional methods for component analysis (e.g. Grassmann averages, PCA, and NMF). The effectiveness of our approach is validated through synthetic data and on the MSRC2 and Sift Flow datasets, demonstrating competitive results in unsupervised and weakly-supervised semantic segmentation.

****

Low-Rank Matrix Factorization Under General Mixture Noise Distributions
Xiangyong Cao, Yang Chen, Qian Zhao, Deyu Meng, Yao Wang, Dong Wang, Zongben Xu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1493-1501

Many computer vision problems can be posed as learning a low-dimensional subspace from high dimensional data. The low rank matrix factorization (LRMF) represents a commonly utilized subspace learning strategy. Most of the current LRMF techniques are constructed on the optimization problem using L_1 norm and L_2 norm, which mainly deal with Laplacian and Gaussian noise, respectively. To make LRMF capable of adapting more complex noise, this paper proposes a new LRMF model by assuming noise as Mixture of Exponential Power (MoEP) distributions and proposes a penalized MoEP model by combining the penalized likelihood method with MoEP distributions. Such setting facilitates the learned LRMF model capable of automatically fitting the real noise through MoEP distributions. Each component in this mixture is adapted from a series of preliminary super- or sub-Gaussian candidates. An Expectation Maximization (EM) algorithm is also designed to infer the parameters involved in the proposed PMoEP model. The advantage of our method is demonstrated by extensive experiments on synthetic data, face modeling and hyperspectral image restoration.

****

Web-Scale Image Clustering Revisited
Yannis Avrithis, Yannis Kalantidis, Evangelos Anagnostopoulos, Ioannis Z. Emiris; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1502-1510

Large scale duplicate detection, clustering and mining of documents or images has been conventionally treated with seed detection via hashing, followed by seed growing heuristics using fast search. Principled clustering methods, especially kernelized and spectral ones, have higher complexity and are difficult to scale above millions. Under the assumption of documents or images embedded in Euclidean space, we revisit recent advances in approximate k-means variants, and borrow their best ingredients to introduce a new one, inverted-quantized k-means (IQ-means). Key underlying concepts are quantization of data points and multi-index based inverted search from centroids to cells. Its quantization is a form of hashing and analogous to seed detection, while its updates are analogous to seed growing, yet principled in the sense of distortion minimization. We further design a dynamic variant that is able to determine the number of clusters k in a single run at nearly zero additional cost. Combined with powerful deep learned representations, we achieve clustering of a 100 million image collection on a single machine in less than one hour.

****

Learning Discriminative Reconstructions for Unsupervised Outlier Removal
Yan Xia, Xudong Cao, Fang Wen, Gang Hua, Jian Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1511-1519

We study the problem of automatically removing outliers from noisy data, with application for removing outlier images from an image collection. We address this problem by utilizing the reconstruction errors of an autoencoder. We observe that when data are reconstructed from low-dimensional representations, the inliers and the outliers can be well separated according to their reconstruction errors. Based on this basic observation, we gradually inject discriminative information in the learning process of an autoencoder to make the inliers and the outliers more separable. Experiments on a variety of image datasets validate our approach.

****

Learning Deconvolution Network for Semantic Segmentation

Hyeonwoo Noh, Seunghoon Hong, Bohyung Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1520-1528

We propose a novel semantic segmentation algorithm by learning a deep deconvolution network. We learn the network on top of the convolutional layers adopted from VGG 16-layer net. The deconvolution network is composed of deconvolution and unpooling layers, which identify pixelwise class labels and predict segmentation masks. We apply the trained network to each proposal in an input image, and construct the final semantic segmentation map by combining the results from all proposals in a simple manner. The proposed algorithm mitigates the limitations of the existing methods based on fully convolutional networks by integrating deep deconvolution network and proposal-wise prediction;our segmentation method typically identifies detailed structures and handles objects in multiple scales naturally. Our network demonstrates outstanding performance in PASCAL VOC 2012 dataset, and we achieve the best accuracy (72.5%) among the methods trained without using Microsoft COCO dataset through ensemble with the fully convolutional network.

****

Conditional Random Fields as Recurrent Neural Networks

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, Philip H. S. Torr; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1529-1537

Pixel-level labelling tasks, such as semantic segmentation, play a central role in image understanding. Recent approaches have attempted to harness the capabilities of deep learning techniques for image recognition to tackle pixel-level labelling tasks. One central issue in this methodology is the limited capacity of deep learning techniques to delineate visual objects. To solve this problem, we introduce a new form of convolutional neural network that combines the strengths of Convolutional Neural Networks (CNNs) and Conditional Random Fields (CRFs)-based probabilistic graphical modelling. To this end, we formulate Conditional Random Fields with Gaussian pairwise potentials and mean-field approximate inference as Recurrent Neural Networks. This network, called CRF-RNN, is then plugged in as a part of a CNN to obtain a deep network that has desirable properties of both CNNs and CRFs. Importantly, our system fully integrates CRF modelling with CNNs, making it possible to train the whole deep network end-to-end with the usual back-propagation algorithm, avoiding offline post-processing methods for object delineation. We apply the proposed method to the problem of semantic image segmentation, obtaining top results on the challenging Pascal VOC 2012 segmentation benchmark.

****

The One Triangle Three Parallelograms Sampling Strategy and Its Application in Shape Regression

Mikael Nilsson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1538-1545

The purpose of this paper is threefold. Firstly, the paper introduces the One Triangle Three Parallelograms (OTTP) sampling strategy, which can be viewed as a way to index pixels from a given shape and image. Secondly, a framework for cascaded shape regression, including the OTTP sampling, is presented. In short, this framework involves binary pixel tests for appearance features combined with shape features followed by a large linear system for each regression stage in the cascade. The proposed solution is found to produce state-of-the-art results on the task of facial landmark estimation. Thirdly, the dependence of accuracy of the landmark predictions and the placement of the mean shape within the detection box is discussed and a method to visualize it is presented.

****

Boosting Object Proposals: From Pascal to COCO

Jordi Pont-Tuset, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1546-1554

Computer vision in general, and object proposals in particular, are nowadays strongly influenced by the databases on which researchers evaluate the performance of their algorithms. This paper studies the transition from the Pascal Visual Object Challenge dataset, which has been the benchmark of reference for the last years, to the updated, bigger, and more challenging Microsoft Common Objects in Context. We first review and deeply analyze the new challenges, and opportunities, that this database presents. We then survey the current state of the art in object proposals and evaluate it focusing on how it generalizes to the new dataset. In sight of these results, we propose various lines of research to take advantage of the new benchmark and improve the techniques. We explore one of these lines, which leads to an improvement over the state of the art of +5.2%.

****

Secrets of GrabCut and Kernel K-Means

Meng Tang, Ismail Ben Ayed, Dmitrii Marin, Yuri Boykov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1555-1563

The log-likelihood energy term in popular model-fitting segmentation methods, e.g. Zhu&Yuille, Chan-Vese, GrabCut, is presented as a generalized "probabilistic K-means" energy for color space clustering. This interpretation reveals some limitations, e.g. over-fitting. We propose an alternative approach to color clustering using kernel K-means energy with well-known properties such as non-linear separation and scalability to higher-dimensional feature spaces. Our bound formulation for kernel K-means allows to combine general pair-wise feature clustering methods with image grid regularization using graph cuts, similarly to standard color model fitting techniques for segmentation. Unlike histogram or GMM fitting, our approach is closely related to average association and normalized cut. But, in contrast to previous pairwise clustering algorithms, our approach can incorporate any standard geometric regularization in the image domain. We analyze extreme cases for kernel bandwidth (e.g. Gini bias) and demonstrate effectiveness of KNN-based adaptive bandwidth strategies. Our kernel K-means approach to segmentation benefits from higher-dimensional features where standard model fitting fails.

****

Video Matting via Sparse and Low-Rank Representation

Dongqing Zou, Xiaowu Chen, Guangying Cao, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1564-1572

We introduce a novel method of video matting via sparse and low-rank representation. Previous matting methods [10, 9] introduced a nonlocal prior to estimate the alpha matte and have achieved impressive results on some data. However, on one hand, searching inadequate or excessive samples may miss good samples or introduce noise; on the other hand, it is difficult to construct consistent nonlocal structures for pixels with similar features, yielding spatially and temporally inconsistent video mattes. In this paper, we proposed a novel video matting method to achieve spatially and temporally consistent matting result. Toward this end, a sparse and low-rank representation model is introduced to pursue consistent nonlocal structures for pixels with similar features. The sparse representation is used to adaptively select best samples and accurately construct the nonlocal structures for all pixels, while the low-rank representation is used to globally ensure consistent nonlocal structures for pixels with similar features. The two representations are combined to generate consistent video mattes. Experimental results show that our method has achieved high quality results in a variety of challenging examples featuring illumination changes, feature ambiguity, topology changes, transparency variation, dis-occlusion, fast motion and motion blur.

****

Joint Object and Part Segmentation Using Deep Learned Potentials

Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, Alan L. Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1573-1581

Segmenting semantic objects from images and parsing them into their respective semantic parts are fundamental steps towards detailed object understanding in com puter vision. In this paper, we propose a joint solution that tackles semantic object and part segmentation simultaneously, in which higher object-level context is provided to guide part segmentation, and more detailed part-level localization is utilized to refine object segmentation. Specifically, we first introduce the concept of semantic compositional parts (SCP) in which similar semantic parts are grouped and shared among different objects. A two-stream fully convolutional network (FCN) is then trained to provide the SCP and object potentials at each pixel. At the same time, a compact set of segments can also be obtained from the SCP predictions of the network. Given the potentials and the generated segments, in order to explore long-range context, we finally construct an efficient fully connected conditional random field (FCRF) to jointly predict the final object and part labels. Extensive evaluation on three different datasets shows that our approach can mutually enhance the performance of object and part segmentation, and outperforms the current state-of-the-art on both tasks.

****

Low-Rank Tensor Constrained Multiview Subspace Clustering

Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, Xiaochun Cao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1582-1590

In this paper, we explore the problem of multiview subspace clustering. We introduce a low-rank tensor constraint to explore the complementary information from multiple views and, accordingly, establish a novel method called Low-rank Tensor constrained Multiview Subspace Clustering (LT-MSC). Our method regards the subspace representation matrices of different views as a tensor, which captures dexterously the high order correlations underlying multiview data. Then the tensor is equipped with a low-rank constraint, which models elegantly the cross information among different views, reduces effectually the redundancy of the learned subspace representations, and improves the accuracy of clustering as well. The inference process of the affinity matrix for clustering is formulated as a tensor nuclear norm minimization problem, constrained with an additional L2,1-norm regularizer and some linear equalities. The minimization problem is convex and thus can be solved efficiently by an Augmented Lagrangian Alternating Direction Minimization (AL-ADM) method. Extensive experimental results on four benchmark datasets show the effectiveness of our proposed LT-MSC method.

****

BodyPrint: Pose Invariant 3D Shape Matching of Human Bodies

Jiangping Wang, Kai Ma, Vivek Kumar Singh, Thomas Huang, Terrence Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1591-1599

3D human body shape matching has large potential on many real world applications, especially with the recent advances in the 3D range sensing technology. We address this problem by proposing a novel holistic human body shape descriptor called BodyPrint. To compute the bodyprint for a given body scan, we fit a deformable human body mesh and project the mesh parameters to a low-dimensional subspace which improves discriminability across different persons. Experiments are carried out on three real-world human body datasets to demonstrate that BodyPrint is robust to pose variation as well as missing information and sensor noise. It improves the matching accuracy significantly compared to conventional 3D shape matching techniques using local features. To facilitate practical applications where the shape database may grow over time, we also extend our learning framework to handle online updates.

****

The Middle Child Problem: Revisiting Parametric Min-Cut and Seeds for Object Proposals

Ahmad Humayun, Fuxin Li, James M. Rehg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1600-1608

Object proposals have recently fueled the progress in detection performance. These proposals aim to provide category-agnostic localizations for all objects in an image. One way to generate proposals is to perform parametric min-cuts over seed locations. This paper demonstrates that standard parametric-cut models are ineffective in obtaining medium-sized objects, which we refer to as the middle chi ld problem. We propose a new energy minimization framework incorporating geodesic distances between segments which solves this problem. In addition, we introduce a new superpixel merging algorithm which can generate a small set of seeds that reliably cover a large number of objects of all sizes. We call our method POISE--- "Proposals for Objects from Improved Seeds and Energies." POISE enables parametric min-cuts to reach their full potential. On PASCAL VOC it generates 2,640 segments with an average overlap of 0.81, whereas the closest competing methods require more than 4,200 proposals to reach the same accuracy. We show detailed quantitative comparisons against 5 state-of-the-art methods on PASCAL VOC and Microsoft COCO segmentation challenges.

****

Contour Guided Hierarchical Model for Shape Matching

Yuanqi Su, Yuehu Liu, Bonan Cuan, Nanning Zheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1609-1617

For its simplicity and effectiveness, star model is popular in shape matching. However, it suffers from the loose geometric connections among parts. In the paper, we present a novel algorithm that reconsiders these connections and reduces the global matching to a set of interrelated local matching. For the purpose, we divide the shape template into overlapped parts and model the matching through a part-based layered structure that uses the latent variable to constrain parts' deformation. As for inference, each part is used for localizing candidates by the partial matching. Thanks to the contour fragments, the partial matching can be solved via modified dynamic programming. The overlapped regions among parts of the template are then explored to make the candidates of parts meet at their shared points. The process is fulfilled via a refined procedure based on iterative dynamic programming. Results on ETHZ shape and Inria Horse datasets demonstrate the benefits of the proposed algorithm.

****

Robust Image Segmentation Using Contour-Guided Color Palettes

Xiang Fu, Chien-Yi Wang, Chen Chen, Changhu Wang, C.-C. Jay Kuo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1618-1625

The contour-guided color palette (CCP) is proposed for robust image segmentation. It efficiently integrates contour and color cues of an image. To find representative colors of an image, color samples along long contours between regions, similar in spirit to machine learning methodology that focus on samples near decision boundaries, are collected followed by the mean-shift (MS) algorithm in the sampled color space to achieve an image-dependent color palette. This color palette provides a preliminary segmentation in the spatial domain, which is further fine-tuned by post-processing techniques such as leakage avoidance, fake boundary removal, and small region mergence. Segmentation performances of CCP and MS are compared and analyzed. While CCP offers an acceptable standalone segmentation result, it can be further integrated into the framework of layered spectral segmentation to produce a more robust segmentation. The superior performance of CCP-based segmentation algorithm is demonstrated by experiments on the Berkeley Segmentation Dataset.

****

Joint Optimization of Segmentation and Color Clustering

Ekaterina Lobacheva, Olga Veksler, Yuri Boykov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1626-1634

Binary energy optimization is a popular approach for segmenting a color image into foreground/background regions. To model the appearance of the regions, color, a relatively high dimensional feature, should be handled effectively. A full color histogram is usually too sparse to be reliable. One approach is to explicitly reduce dimensionality by clustering or quantizing the color space. Another popular approach is to fit GMMs for soft implicit clustering of the color space. These approaches work well when the foreground/background are sufficiently distinct. In cases of more subtle difference in appearance, both approaches may reduce or even eliminate foreground/background distinction. This happens because either color clustering is performed completely independently from the segmentation process, as a preprocessing step (in clustering), or independently for the foregro und and independently for the background (in GMM). We propose to make clustering an integral part of segmentation, by including a new clustering term in the energy function. Our energy function with a clustering term favours clusterings that make foreground/background appearance more distinct. Thus our energy function jointly optimizes over color clustering, foreground/background models, and segmentation. Exact optimization is not feasible, therefore we develop an approximate algorithm. We show the advantage of including the color clustering term into the energy function on camouflage images, as well as standard segmentation datasets.

****

BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation

Jifeng Dai, Kaiming He, Jian Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1635-1643

Recent leading approaches to semantic segmentation rely on deep convolutional networks trained with human-annotated, pixel-level segmentation masks. Such pixel-accurate supervision demands expensive labeling effort and limits the performance of deep networks that usually benefit from more training data. In this paper, we propose a method that achieves competitive accuracy but only requires easily obtained bounding box annotations. The basic idea is to iterate between automatically generating region proposals and training convolutional networks. These two steps gradually recover segmentation masks for improving the networks, and vise versa. Our method, called "BoxSup", produces competitive results (e.g., 62.0% mAP for validation) supervised by boxes only, on par with strong baselines (e.g., 63.8% mAP) fully supervised by masks under the same setting. By leveraging a large amount of bounding boxes, BoxSup further yields state-of-the-art results on PASCAL VOC 2012 and PASCAL-CONTEXT.

****

Detection and Segmentation of 2D Curved Reflection Symmetric Structures

Ching L. Teo, Cornelia Fermuller, Yiannis Aloimonos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1644-1652

Symmetry, as one of the key components of Gestalt theory, provides an important mid-level cue that serves as input to higher visual processes such as segmentation. In this work, we propose a complete approach that links the detection of curved reflection symmetries to produce symmetry-constrained segments of structures/regions in real images with clutter. For curved reflection symmetry detection, we leverage on patch-based symmetric features to train a Structured Random Forest classifier that detects multiscaled curved symmetries in 2D images. Next, using these curved symmetries, we modulate a novel symmetry-constrained foreground-background segmentation by their symmetry scores so that we enforce global symmetrical consistency in the final segmentation. This is achieved by imposing a pairwise symmetry prior that encourages symmetric pixels to have the same labels over a MRF-based representation of the input image edges, and the final segmentation is obtained via graph-cuts. Experimental results over four publicly available datasets containing annotated symmetric structures: 1) SYMMAX-300, 2) BSD-Parts, 3) Weizmann Horse and 4) NY-roads demonstrate the approach's applicability to different environments with state-of-the-art performance.

****

Unsupervised Tube Extraction Using Transductive Learning and Dense Trajectories

Mihai Marian Puscas, Enver Sangineto, Dubravko Culibrk, Nicu Sebe; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1653-1661

We address the problem of automatic extraction of foreground objects from videos. The goal is to provide a method for unsupervised collection of samples which can be further used for object detection training without any human intervention. We use the well known Selective Search approach to produce an initial still-image based segmentation of the video frames.   This initial set of proposals is pruned and temporally extended using optical flow and transductive learning.   Specifically, we propose to use Dense Trajectories in order to robustly match and track candidate boxes over different frames. The obtained box tracks are used to collect samples for unsupervised training of track-specific detectors. Finally, the detectors are run on the videos to extract the final tubes. The combination of appearance-based static ''objectness'' (Selective Search), motion information (Dense Trajectories) and transductive learning (detectors are forced to "overfit" on the unsupervised data used for training) makes the proposed approach extremely robust. We outperform state-of-the-art systems by a large margin on common benchmarks used for tube proposal evaluation.

****

Compositional Hierarchical Representation of Shape Manifolds for Classification of Non-Manifold Shapes

Mete Ozay, Umit Rusen Aktas, Jeremy L. Wyatt, Ales Leonardis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1662-1670

We address the problem of statistical learning of shape models which are invariant to translation, rotation and scale in compositional hierarchies when data spaces of measurements and shape spaces are not topological manifolds. In practice, this problem is observed while modeling shapes having multiple disconnected components, e.g. partially occluded shapes in cluttered scenes. We resolve the aforementioned problem by first reformulating the relationship between data and shape spaces considering the interaction between Receptive Fields (RFs) and Shape Manifolds (SMs) in a compositional hierarchical shape vocabulary. Then, we suggest a method to model the topological structure of the SMs for statistical learning of the geometric transformations of the shapes that are defined by group actions on the SMs. For this purpose, we design a disjoint union topology using an indexing mechanism for the formation of shape models on SMs in the vocabulary, recursively. We represent the topological relationship between shape components using graphs, which are aggregated to construct a hierarchical graph structure for the shape vocabulary. To this end, we introduce a framework to implement the indexing mechanisms for the employment of the vocabulary for structural shape classification. The proposed approach is used to construct invariant shape representations. Results on benchmark shape classification outperform state-of-the-art methods.

****

Shell PCA: Statistical Shape Modelling in Shell Space

Chao Zhang, Behrend Heeren, Martin Rumpf, William A. P. Smith; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1671-1679

In this paper we describe how to perform Principal Components Analysis in "shell space". Thin shells are a physical model for surfaces with non-zero thickness whose deformation dissipates elastic energy. Thin shells, or their discrete counterparts, can be considered to reside in a shell space in which the notion of distance is given by the elastic energy required to deform one shape into another. It is in this setting that we show how to perform statistical analysis of a set of shapes (meshes in dense correspondence), providing a hybrid between physical and statistical shape modelling. The resulting models are better able to capture non-linear deformations, for example resulting from articulated motion, even when training data is very sparse compared to the dimensionality of the observation space.

****

Learning to Combine Mid-Level Cues for Object Proposal Generation

Tom Lee, Sanja Fidler, Sven Dickinson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1680-1688

In recent years, region proposals have replaced sliding windows in support of object recognition, offering more discriminating shape and appearance information through improved localization. One powerful approach for generating region proposals is based on minimizing parametric energy functions with parametric maxflow. In this paper, we introduce Parametric Min-Loss (PML), a novel structured learning framework for parametric energy functions. While PML is generally applicable to different domains, we use it in the context of region proposals to learn to combine a set of mid-level grouping cues to yield a small set of object region proposals with high recall. Our learning framework accounts for multiple diverse outputs, and is complemented by diversification seeds based on image locatio n and color. This approach casts perceptual grouping and cue combination in a novel structured learning framework which yields baseline improvements on VOC 2012 and COCO 2014.

****

Enhancing Road Maps by Parsing Aerial Images Around the World

Gellert Mattyus, Shenlong Wang, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1689-1697

In recent years, contextual models that exploit maps have been shown to be very effective for many recognition and localization tasks. In this paper we propose to exploit aerial images in order to enhance freely available world maps. Towards this goal, we make use of OpenStreetMap and formulate the problem as the one of inference in a Markov random field parameterized in terms of the location of the road-segment centerlines as well as their width. This parameterization enables very efficient inference and returns only topologically correct roads. In particular, we can segment all OSM roads in the whole world in a single day using a small cluster of 10 computers. Importantly, our approach generalizes very well; it can be trained using only 1.5 km2 aerial imagery and produce very accurate results in any location across the globe. We demonstrate the effectiveness of our approach outperforming the state-of-the-art in two new benchmarks that we collect. We then show how our enhanced maps are beneficial for semantic segmentation of ground images.

****

Probabilistic Appearance Models for Segmentation and Classification

Julia Kruger, Jan Ehrhardt, Heinz Handels; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1698-1706

Statistical shape and appearance models are often based on the accurate identification of one-to-one correspondences in a training data set. At the same time, the determination of these corresponding landmarks is the most challenging part of such methods. Hufnagel etal developed an alternative method using correspondence probabilities for a statistical shape model. We propose the use of probabilistic correspondences for statistical appearance models by incorporating appearance information into the framework. A point-based representation is employed representing the image by a set of vectors assembling position and appearances. Using probabilistic correspondences between these multi-dimensional feature vectors eliminates the need for extensive preprocessing to find corresponding landmarks and reduces the dependence of the generated model on the landmark positions. Then, a maximum a-posteriori approach is used to derive a single global optimization criterion with respect to model parameters and observation dependent parameters, that directly affects shape and appearance information of the considered structures. Model generation and fitting can be expressed by optimizing the same criterion. The developed framework describes the modeling process in a concise and flexible mathematical way and allows for additional constraints as topological regularity in the modeling process. Furthermore, it eliminates the demand for costly correspondence determination. We apply the model for segmentation and landmark identification in hand X-ray images, where segmentation information is modeled as further features in the vectorial image representation. The results demonstrate the feasibility of the model to reconstruct contours and landmarks for unseen test images. Furthermore, we apply the model for tissue classification, where a model is generated for healthy brain tissue using 2D MRI slices. Applying the model to images of stroke patients the probabilistic correspondences are used to classify between healthy and pathological structures. The results demonstrate the ability of the probabilistic model to recognize healthy and pathological tissue automatically.

****

A Randomized Ensemble Approach to Industrial CT Segmentation

Hyojin Kim, Jayaraman Jayaraman J. Thiagarajan, Peer-Timo Bremer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1707-1715

Tuning the models and parameters of common segmentation approaches is challenging especially in the presence of noise and artifacts. Ensemble-based techniques a ttempt to compensate by randomly varying models and/or parameters to create a diverse set of hypotheses, which are subsequently ranked to arrive at the best solution. However, these methods have been restricted to cases where the underlying models are well-established, e.g. natural images. In practice, it is difficult to determine a suitable base-model and the amount of randomization required. Furthermore, for multi-object scenes no single hypothesis may perform well for all objects, reducing the overall quality of the results. This paper presents a new ensemble-based segmentation framework for industrial CT images demonstrating that comparatively simple models and randomization strategies can significantly improve the result over existing techniques. Furthermore, we introduce a per-object based ranking, followed by a consensus inference that can outperform even the best case scenario of existing hypothesis ranking approaches. We demonstrate the effectiveness of our approach using a set of noise and artifact rich CT images from baggage security and show that it significantly outperforms existing solutions in this area.

****

Semi-Supervised Normalized Cuts for Image Segmentation

Selene E. Chew, Nathan D. Cahill; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1716-1723

Since its introduction as a powerful graph-based method for image segmentation, the Normalized Cuts (NCuts) algorithm has been generalized to incorporate expert knowledge about how certain pixels or regions should be grouped, or how the resulting segmentation should be biased to be correlated with priors. Previous approaches incorporate hard must-link constraints on how certain pixels should be grouped as well as hard cannot-link constraints on how other pixels should be separated into different groups. In this paper, we reformulate NCuts to allow both sets of constraints to be handled in a soft manner, enabling the user to tune the degree to which the constraints are satisfied. An approximate spectral solution to the reformulated problem exists without requiring explicit construction of a large, dense matrix; hence, computation time is comparable to that of unconstrained NCuts. Using synthetic data and real imagery, we show that soft handling of constraints yields better results than unconstrained NCuts and enables more robust clustering and segmentation than is possible when the constraints are strictly enforced.

****

StereoSnakes: Contour Based Consistent Object Extraction For Stereo Images

Ran Ju, Tongwei Ren, Gangshan Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1724-1732

Consistent object extraction plays an essential role for stereo image editing with the population of stereoscopic 3D media. Most previous methods perform segmentation on entire images for both views using dense stereo correspondence constraints. We find that for such kind of methods the computation is highly redundant since the two views are near-duplicate. Besides, the consistency may be violated due to the imperfectness of current stereo matching algorithms. In this paper, we propose a contour based method which searches for consistent object contours instead of regions. It integrates both stereo correspondence and object boundary constraints into an energy minimization framework. The proposed method has several advantages compared to previous works. First, the searching space is restricted in object boundaries thus the efficiency significantly improved. Second, the discriminative power of object contours results in a more consistent segmentation. Furthermore, the proposed method can effortlessly extend existing single-image segmentation methods to work in stereo scenarios. The experiment on the Adobe benchmark shows superior extraction accuracy and significant improvement of efficiency of our method to state-of-the-art. We also demonstrate in a few applications how our method can be used as a basic tool for stereo image editing.

****

Semantic Segmentation of RGBD Images With Mutex Constraints

Zhuo Deng, Sinisa Todorovic, Longin Jan Latecki; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1733-1741

In this paper, we address the problem of semantic scene segmentation of RGB-D im ages of indoor scenes. We propose a novel image region labeling method which augments CRF formulation with hard mutual exclusion (mutex) constraints. This way our approach can make use of rich and accurate 3D geometric structure coming from Kinect in a principled manner. The final labeling result must satisfy all mutex constraints, which allows us to eliminate configurations that violate common sense physics laws like placing a floor above a night stand. Three classes of mutex constraints are proposed: global object co-occurrence constraint, relative height relationship constraint, and local support relationship constraint. We evaluate our approach on the NYU-Depth V2 dataset, which consists of 1449 cluttered indoor scenes, and also test generalization of our model trained on NYU-Depth V2 dataset directly on a recent SUN3D dataset without any new training. The experimental results show that we significantly outperform the state-of-the-art methods in scene labeling on both datasets.

****

Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation

George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, Alan L. Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1742-1750

Deep convolutional neural networks (DCNNs) trained on a large number of images with strong pixel-level annotations have recently significantly pushed the state-of-art in semantic image segmentation. We study the more challenging problem of learning DCNNs for semantic image segmentation from either (1) weakly annotated training data such as bounding boxes or image-level labels or (2) a combination of few strongly labeled and many weakly labeled images, sourced from one or multiple datasets. We develop Expectation-Maximization (EM) methods for semantic image segmentation model training under these weakly supervised and semi-supervised settings. Extensive experimental evaluation shows that the proposed techniques can learn models delivering competitive results on the challenging PASCAL VOC 2012 image segmentation benchmark, while requiring significantly less annotation effort. We share source code implementing the proposed system at https://bitbucket.org/deeplab/deeplab-public.

****

Efficient Decomposition of Image and Mesh Graphs by Lifted Multicuts

Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoue, Thomas Brox, Bjorn Andres; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1751-1759

Formulations of the Image Decomposition Problem as a Multicut Problem (MP) w.r.t. a superpixel graph have received considerable attention. In contrast, instances of the MP w.r.t. a pixel grid graph have received little attention, firstly, because the MP is NP-hard and instances w.r.t. a pixel grid graph are hard to solve in practice, and, secondly, due to the lack of long-range terms in the objective function of the MP. We propose a generalization of the MP with long-range terms (LMP). We design and implement two efficient algorithms (primal feasible heuristics) for the MP and LMP which allow us to study instances of both problems w.r.t. the pixel grid graphs of the images in the BSDS-500 benchmark. The decompositions we obtain do not differ significantly from the state of the art, suggesting that the LMP is a competitive formulation of the Image Decomposition Problem. To demonstrate the generality of the LMP, we apply it also to the Mesh Decomposition Problem posed by the Princeton benchmark, obtaining state-of-the-art decompositions.

****

Parsimonious Labeling

Puneet K. Dokania, M. Pawan Kumar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1760-1768

We propose a new family of discrete energy minimization problems, which we call parsimonious labeling. Our energy function consists of unary potentials and high-order clique potentials. While the unary potentials are arbitrary, the clique potentials are proportional to the diversity of the set of unique labels assigned to the clique. Intuitively, our energy function encourages the labeling to be p

arsimonious, that is, use as few labels as possible. This in turn allows us to capture useful cues for important computer vision applications such as stereo correspondence and image denoising. Furthermore, we propose an efficient graph-cuts based algorithm for the parsimonious labeling problem that provides strong theoretical guarantees on the quality of the solution. Our algorithm consists of three steps. First, we approximate a given diversity using a mixture of a novel hierarchical Pn Potts model. Second, we use a divide-and-conquer approach for each mixture component, where each subproblem is solved using an efficient alpha-expansion algorithm. This provides us with a small number of putative labelings, one for each mixture component. Third, we choose the best putative labeling in terms of the energy value. Using both synthetic and standard real datasets, we show that our algorithm significantly outperforms other graph-cuts based approaches.
****

Volumetric Bias in Segmentation and Reconstruction: Secrets and Solutions
Yuri Boykov, Hossam Isack, Carl Olsson, Ismail Ben Ayed; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1769-1777
Many standard optimization methods for segmentation and reconstruction compute ML model estimates for appearance or geometry of segments, e.g. Zhu-Yuille 1996, Torr 1998, Chan-Vese 2001, GrabCut 2004, Delong et al. 2012. We observe that the standard likelihood term in these formulations corresponds to a generalized probabilistic K-means energy. In learning it is well known that this energy has a strong bias to clusters of equal size, which we express as a penalty for KL divergence from a uniform distribution of cardinalities. However, this volumetric bias has been mostly ignored in computer vision. We demonstrate significant artifacts in standard segmentation and reconstruction methods due to this bias. Moreover, we propose binary and multi-label optimization techniques that either (a) remove this bias or (b) replace it by a KL divergence term for any given target volume distribution. Our general ideas apply to continuous or discrete energy formulations in segmentation, stereo, and other reconstruction problems.
****

Entropy Minimization for Convex Relaxation Approaches
Mohamed Souiai, Martin R. Oswald, Youngwook Kee, Junmo Kim, Marc Pollefeys, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1778-1786
Despite their enormous success in solving hard combinatorial problems, convex relaxation approaches often suffer from the fact that the computed solutions are far from binary and that subsequent heuristic binarization may substantially degrade the quality of computed solutions. In this paper, we propose a novel relaxation technique which incorporates the entropy of the objective variable as a measure of relaxation tightness. We show both theoretically and experimentally that augmenting the objective function with an entropy term gives rise to more binary solutions and consequently solutions with a substantially tighter optimality gap. We use difference of convex function (DC) programming as an efficient and provably convergent solver for the arising convex-concave minimization problem. We evaluate this approach on three prominent non-convex computer vision challenges: multi-label inpainting, image segmentation and spatio-temporal multi-view reconstruction. These experiments show that our approach consistently yields better solutions with respect to the original integral optimization problem
****

Adaptively Unified Semi-Supervised Dictionary Learning With Active Points
Xiaobo Wang, Xiaojie Guo, Stan Z. Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1787-1795
Semi-supervised dictionary learning aims to construct a dictionary by utilizing both labeled and unlabeled data. To enhance the discriminative capability of the learned dictionary, numerous discriminative terms have been proposed by evaluating either the prediction loss or the class separation criterion on the coding vectors of labeled data, but with rare consideration of the power of the coding vectors corresponding to unlabeled data. In this paper, we present a novel semi-supervised dictionary learning method, which uses the informative coding vectors of both labeled and unlabeled data, and adaptively emphasizes the high confidenc

e coding vectors of unlabeled data to enhance the dictionary discriminative capability simultaneously. By doing so, we integrate the discrimination of dictionary, the induction of classifier to new testing data and the transduction of labels to unlabeled data into a unified framework. To solve the proposed problem, an effective iterative algorithm is designed. Experimental results on a series of benchmark databases show that our method outperforms other state-of-the-art dictionary learning methods in most cases.
****

Constrained Convolutional Neural Networks for Weakly Supervised Segmentation
Deepak Pathak, Philipp Krahenbuhl, Trevor Darrell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1796-1804

We present an approach to learn a dense pixel-wise labeling from image-level tags. Each image-level tag imposes constraints on the output labeling of a Convolutional Neural Network (CNN) classifier. We propose Constrained CNN (CCNN), a method which uses a novel loss function to optimize for any set of linear constraints on the output space (i.e. predicted label distribution) of a CNN. Our loss formulation is easy to optimize and can be incorporated directly into standard stochastic gradient descent optimization. The key idea is to phrase the training objective as a biconvex optimization for linear models, which we then relax to nonlinear deep networks. Extensive experiments demonstrate the generality of our new learning framework. The constrained loss yields state-of-the-art results on weakly supervised semantic image segmentation. We further demonstrate that adding slightly more supervision can greatly improve the performance of the learning algorithm.
****

A Multiscale Variable-Grouping Framework for MRF Energy Minimization
Omer Meir, Meirav Galun, Stav Yagev, Ronen Basri, Irad Yavneh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1805-1813

We present a multiscale approach for minimizing the energy associated with Markov Random Fields (MRFs) with energy functions that include arbitrary pairwise potentials. The MRF is represented on a hierarchy of successively coarser scales, where the problem on each scale is itself an MRF with suitably defined potentials. These representations are used to construct an efficient multiscale algorithm that seeks a minimal-energy solution to the original problem. The algorithm is iterative and features a bidirectional crosstalk between fine and coarse representations. We use consistency criteria to guarantee that the energy is nonincreasing throughout the iterative process. The algorithm is evaluated on real-world datasets, achieving competitive performance in relatively short run-times.
****

Inferring M-Best Diverse Labelings in a Single One
Alexander Kirillov, Bogdan Savchynskyy, Dmitrij Schlesinger, Dmitry Vetrov, Carsten Rother; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1814-1822

We consider the task of finding M-best diverse solutions in a graphical model. In a previous work by Batra et al. an algorithmic approach for finding such solutions was proposed, and its usefulness was shown in numerous applications. Contrary to previous work we propose a novel formulation of the problem in form of a single energy minimization problem in a specially constructed graphical model. We show that the method of Batra et al. can be considered as a greedy approximate algorithm for our model, whereas we introduce an efficient specialized optimization technique for it, based on alpha-expansion. We evaluate our method on two application scenarios, interactive and semantic image segmentation, with binary and multiple labels. In both cases we achieve considerably better error rates than state-of-the art diversity methods. Furthermore, we empirically discover that in the binary label case we were able to reach global optimality for all test instances.
****

Convolutional Sparse Coding for Image Super-Resolution
Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 182

3-1831
Sparse coding (SC) plays an important role in versatile computer vision applications such as image super-resolution (SR). Most of the previous SC based SR methods partition the image into overlapped patches, and process each patch separately. These methods, however, ignore the consistency of pixels in overlapped patches, which is a strong constraint for image reconstruction. In this paper, we propose a convolutional sparse coding (CSC) based SR (CSC-SR) method to address the consistency issue. Our CSC-SR involves three groups of parameters to be learned: (i) a set of filters to decompose the low resolution (LR) image into LR sparse feature maps; (ii) a mapping function to predict the high resolution (HR) feature maps from the LR ones; and (iii) a set of filters to reconstruct the HR images from the predicted HR feature maps via simple convolution operations. By working directly on the whole image, the proposed CSC-SR algorithm does not need to divide the image into overlapped patches, and can exploit the image global correlation to produce more robust reconstruction of image local structures. Experimental results clearly validate the advantages of CSC over patch based SC in SR application. Compared with state-of-the-art SR methods, the proposed CSC-SR method achieves highly competitive PSNR results, while demonstrating better edge and texture preservation performance.
****
A Wavefront Marching Method for Solving the Eikonal Equation on Cartesian Grids
Brais Cancela, Marcos Ortega, Manuel G. Penedo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1832-1840
This paper presents a new wavefront propagation method for dealing with the classic Eikonal equation. While classic Dijkstra-like graph-based techniques achieve the solution in O(M log M), they do not approximate the unique physically relevant solution very well. Fast Marching Methods (FMM) were created to efficiently solve the continuous problem. The proposed approximation tries to maintain the complexity, in order to make the algorithm useful in a wide range of contexts. The key idea behind our method is the creation of 'mini wave-fronts', which are combined to propagate the solution. Experimental results show the improvement in the accuracy with respect to the state of the art, while the average computational speed is maintained in O(M log M), similar to the FMM techniques.
****
A Projection Free Method for Generalized Eigenvalue Problem With a Nonsmooth Regularizer
Seong Jae Hwang, Maxwell D. Collins, Sathya N. Ravi, Vamsi K. Ithapu, Nagesh Adluru, Sterling C. Johnson, Vikas Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1841-1849
Eigenvalue problems are ubiquitous in computer vision, covering a very broad spectrum of applications ranging from estimation problems in multi-view geometry to image segmentation. Few other linear algebra problems have a more mature set of numerical routines available and many computer vision libraries leverage such tools extensively. However, the ability to call the underlying solver only as a ``black box'' can often become restrictive. Many `human in the loop' settings in vision frequently exploit supervision from an expert, to the extent that the user can be considered a subroutine in the overall system. In other cases, there is additional domain knowledge, side or even partial information that one may want to incorporate within the formulation. In general, regularizing a (generalized) eigenvalue problem with such side information remains difficult. Motivated by these needs, this paper presents an optimization scheme to solve generalized eigenvalue problems (GEP) involving a (nonsmooth) regularizer. We start from an alternative formulation of GEP where the feasibility set of the model involves the Stiefel manifold. The core of this paper presents an end to end stochastic optimization scheme for the resultant problem. We show how this general algorithm enables improved statistical analysis of brain imaging data where the regularizer is derived from other `views' of the disease pathology, involving clinical measurements and other image-derived representations.
****
Optimizing Expected Intersection-Over-Union With Candidate-Constrained CRFs

Faruk Ahmed, Dany Tarlow, Dhruv Batra; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1850-1858

We study the question of how to make loss-aware predictions in image segmentation settings where the evaluation function is the Intersection-over-Union (IoU) measure that is used widely in evaluating image segmentation systems. Currently, there are two dominant approaches: the first approximates the Expected-IoU (EIoU) score as Expected-Intersection-over-Expected-Union (EIoEU); and the second approach is to compute exact EIoU but only over a small set of high-quality candidate solutions. We begin by asking which approach we should favor for two typical image segmentation tasks. Studying this question leads to two new methods that draw ideas from both existing approaches. Our new methods use the EIoEU approximation paired with high quality candidate solutions. Experimentally we show that our new approaches lead to improved performance on both image segmentation tasks.

****

Higher-Order Inference for Multi-Class Log-Supermodular Models

Jian Zhang, Josip Djolonga, Andreas Krause; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1859-1867

Higher-order models have been shown to be very useful for a plethora of computer vision tasks. However, existing techniques have focused mainly on MAP inference. In this paper, we present the first efficient approach towards approximate Bayesian marginal inference in a general class of high-order, multi-label attractive models, where previous techniques slow down exponentially with the order (clique size). We formalize this task as performing inference in log-supermodular models under partition constraints, and present an efficient variational inference technique. The resulting optimization problems are convex and yield bounds on the partition function. We also obtain a fully factorized approximation to the posterior, which can be used in lieu of the true complicated distribution. We empirically demonstrate the performance of our approach by comparing it to traditional inference methods on a challenging high-fidelity multi-label image segmentation dataset. We obtain state-of-the-art classification accuracy for MAP inference, and substantially improved ROC curves using the approximate marginals.

****

Depth-Based Hand Pose Estimation: Data, Methods, and Challenges

James S. Supancic III, Gregory Rogez, Yi Yang, Jamie Shotton, Deva Ramanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1868-1876

Hand pose estimation has matured rapidly in recent years. The introduction of commodity depth sensors and a multitude of practical applications have spurred new advances. We provide an extensive analysis of the state-of-the-art, focusing on hand pose estimation from a single depth frame. To do so, we have implemented a considerable number of systems, and will release all software and evaluation code. We summarize important conclusions here: (1) Pose estimation appears roughly solved for scenes with isolated hands. However, methods still struggle to analyze cluttered scenes where hands may be interacting with nearby objects and surfaces. To spur further progress we introduce a challenging new dataset with diverse, cluttered scenes. (2) Many methods evaluate themselves with disparate criteria, making comparisons difficult. We define a consistent evaluation criteria, rigorously motivated by human experiments. (3) We introduce a simple nearest-neighbor baseline that outperforms most existing systems. This implies that most systems do not generalize beyond their training sets. This also reinforces the under-appreciated point that training data is as important as the model itself. We conclude with directions for future progress.

****

Adaptive Dither Voting for Robust Spatial Verification

Xiaomeng Wu, Kunio Kashino; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1877-1885

Hough voting in a geometric transformation space allows us to realize spatial verification, but remains sensitive to feature detection errors because of the inflexible quantization of single feature correspondences. To handle this problem, we propose a new method, called adaptive dither voting, for robust spatial verif ication. For each correspondence, instead of hard-mapping it to a single transformation, the method augments its description by using multiple dithered transformations that are deterministically generated by the other correspondences. The method reduces the probability of losing correspondences during transformation quantization, and provides high robustness as regards mismatches by imposing three geometric constraints on the dithering process. We also propose exploiting the non-uniformity of a Hough histogram as the spatial similarity to handle multiple matching surfaces. Extensive experiments conducted on four datasets show the superiority of our method. The method outperforms its state-of-the-art counterparts in both accuracy and scalability, especially when it comes to the retrieval of small, rotated objects.

****

Alternating Co-Quantization for Cross-Modal Hashing

Go Irie, Hiroyuki Arai, Yukinobu Taniguchi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1886-1894

This paper addresses the problem of unsupervised learning of binary hash codes for efficient cross-modal retrieval. Many unimodal hashing studies have proven that both similarity preservation of data and maintenance of quantization quality are essential for improving retrieval performance with binary hash codes. However, most existing cross-modal hashing methods mainly have focused on the former, and the latter still remains almost untouched. We propose a method to minimize the binary quantization errors, which is tailored to cross-modal hashing. Our approach, named Alternating Co-Quantization (ACQ), alternately seeks binary quantizers for each modality space with the help of connections to other modality data so that they give minimal quantization errors while preserving data similarities. ACQ can be coupled with various existing cross-modal dimension reduction methods such as Canonical Correlation Analysis (CCA) and substantially boosts their retrieval performance in the Hamming space. Extensive experiments demonstrate that ACQ can outperform several state-of-the-art methods, even when it is combined with simple CCA.

****

Learning Deep Representation With Large-Scale Attributes

Wanli Ouyang, Hongyang Li, Xingyu Zeng, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1895-1903

Learning strong feature representations from large scale supervision has achieved remarkable success in computer vision as the emergence of deep learning techniques. It is driven by big visual data with rich annotations. This paper contributes a large-scale object attribute database (The dataset is available on www.ee.cuhk.edu.hk/ xgwang/ImageNetAttribute.html) that contains rich attribute annotations (over 300 attributes) for ~180k samples and 494 object classes. Based on the ImageNet object detection dataset, it annotates the rotation, viewpoint, object part location, part occlusion, part existence, common attributes, and class-specific attributes. Then we use this dataset to train deep representations and extensively evaluate how these attributes are useful on the general object detection task. In order to make better use of the attribute annotations, a deep learning scheme is proposed by modeling the relationship of attributes and hierarchically clustering them into semantically meaningful mixture types. Experimental results show that the attributes are helpful in learning better features and improving the object detection accuracy by 2.6% in mAP on the ILSVRC 2014 object detection dataset and 2.4% in mAP on PASCAL VOC 2007 object detection dataset. Such improvement is well generalized across datasets.

****

Deep Learning Strong Parts for Pedestrian Detection

Yonglong Tian, Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1904-1912

Recent advances in pedestrian detection are attained by transferring the learned features of Convolutional Neural Network (ConvNet) to pedestrians. This ConvNet is typically pre-trained with massive general object categories (e.g. ImageNet). Although these features are able to handle variations such as poses, viewpoints, and lightings, they may fail when pedestrian images with complex occlusions a re present. Occlusion handling is one of the most important problem in pedestrian detection. Unlike previous deep models that directly learned a single detector for pedestrian detection, we propose DeepParts, which consists of extensive part detectors. DeepParts has several appealing properties. First, DeepParts can be trained on weakly labeled data, i.e. only pedestrian bounding boxes without part annotations are provided. Second, DeepParts is able to handle low IoU positive proposals that shift away from ground truth. Third, each part detector in DeepParts is a strong detector that can detect pedestrian by observing only a part of a proposal. Extensive experiments in Caltech dataset demonstrate the effectiveness of DeepParts, which yields a new state-of-the-art miss rate of 11:89%, outperforming the second best method by 10%.

****

Flowing ConvNets for Human Pose Estimation in Videos

Tomas Pfister, James Charles, Andrew Zisserman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1913-1921

The objective of this work is human pose estimation in videos, where multiple frames are available. We investigate a ConvNet architecture that is able to benefit from temporal context by combining information across the multiple frames using optical flow. To this end we propose a network architecture with the following novelties: (i) a deeper network than previously investigated for regressing heatmaps; (ii) spatial fusion layers that learn an implicit spatial model; (iii) optical flow is used to align heatmap predictions from neighbouring frames; and (iv) a final parametric pooling layer which learns to combine the aligned heatmaps into a pooled confidence map. We show that this architecture outperforms a number of others, including one that uses optical flow solely at the input layers, one that regresses joint coordinates directly, and one that predicts heatmaps without spatial fusion. The new architecture outperforms the state of the art by a large margin on three video pose estimation datasets, including the very challenging Poses in the Wild dataset, and outperforms other deep methods that don't use a graphical model on the single-image FLIC benchmark (and also Chen & Yuille and Tompson et al. in the high precision region).

****

Top Rank Supervised Binary Coding for Visual Search

Dongjin Song, Wei Liu, Rongrong Ji, David A. Meyer, John R. Smith; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1922-1930

In recent years, binary coding techniques are becoming increasingly popular because of their high efficiency in handling large-scale computer vision applications. It has been demonstrated that supervised binary coding techniques that leverage supervised information can significantly enhance the coding quality, and hence greatly benefit visual search tasks. Typically, a modern binary coding method seeks to learn a group of coding functions which compress data samples into binary codes. However, few methods pursued the coding functions such that the precision at the top of a ranking list according to Hamming distances of the generated binary codes is optimized. In this paper, we propose a novel supervised binary coding approach, namely Top Rank Supervised Binary Coding (Top-RSBC), which explicitly focuses on optimizing the precision of top positions in a Hamming-distance ranking list towards preserving the supervision information. The core idea is to train the disciplined coding functions, by which the mistakes at the top of a Hamming-distance ranking list are penalized more than those at the bottom. To solve such coding functions, we relax the original discrete optimization objective with a continuous surrogate, and derive a stochastic gradient descent to optimize the surrogate objective. To further reduce the training time cost, we also design an online learning algorithm to optimize the surrogate objective more efficiently. Empirical studies based upon three benchmark image datasets demonstrate that the proposed binary coding approach achieves superior image search accuracy over the state-of-the-arts.

****

BubbLeNet: Foveated Imaging for Visual Discovery

Kevin Matzen, Noah Snavely; Proceedings of the IEEE International Conference on

Computer Vision (ICCV), 2015, pp. 1931-1939

We propose a new method for turning an Internet-scale corpus of categorized images into a small set of human-interpretable discriminative visual elements using powerful tools based on deep learning. A key challenge with deep learning methods is generating human-interpretable models. To address this, we propose a new technique that uses bubble images -- images where most of the content has been obscured -- to identify spatially localized, discriminative content in each image. By modifying the model training procedure to use both the source imagery and these bubble images, we can arrive at final models which retain much of the original classification performance, but are much more amenable to identifying interpretable visual elements. We apply our algorithm to a wide variety of datasets, including two new Internet-scale datasets of people and places, and show applications to visual mining and discovery. Our method is simple, scalable, and produces visual elements that are highly representative compared to prior work.

****

PQTable: Fast Exact Asymmetric Distance Neighbor Search for Product Quantization Using Hash Tables

Yusuke Matsui, Toshihiko Yamasaki, Kiyoharu Aizawa; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1940-1948

We propose the product quantization table (PQTable), a product quantization-based hash table that is fast and requires neither parameter tuning nor training steps. The PQTable produces exactly the same results as a linear PQ search, and is 10^2 to 10^5 times faster when tested on the SIFT1B data. In addition, although state-of-the-art performance can be achieved by previous inverted-indexing-based approaches, such methods do require manually designed parameter setting and much training, whereas our method is free from them. Therefore, PQTable offers a practical and useful solution for real-world problems.

****

Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions

Sven Bambach, Stefan Lee, David J. Crandall, Chen Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1949-1957

Hands appear very often in egocentric video, and their appearance and pose give important cues about what people are doing and what they are paying attention to. But existing work in hand detection has made strong assumptions that work well in only simple scenarios, such as with limited interaction with other people or in lab settings. We develop methods to locate and distinguish between hands in egocentric video using strong appearance models with Convolutional Neural Networks, and introduce a simple candidate region generation approach that outperforms existing techniques at a fraction of the computational cost. We show how these high-quality bounding boxes can be used to create accurate pixelwise hand regions, and as an application, we investigate the extent to which hand segmentation alone can distinguish between different activities. We evaluate these techniques on a new dataset of 48 first-person videos (along with pixel-level ground truth for over 15,000 hand instances) of people interacting in realistic environments.

****

Fast and Accurate Head Pose Estimation via Random Projection Forests

Donghoon Lee, Ming-Hsuan Yang, Songhwai Oh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1958-1966

In this paper, we consider the problem of estimating the gaze direction of a person from a low-resolution image. Under this condition, reliably extracting facial features is very difficult. We propose a novel head pose estimation algorithm based on compressive sensing. Head image patches are mapped to a large feature space using the proposed extensive, yet efficient filter bank. The filter bank is designed to generate sparse responses of color and gradient information, which can be compressed using random projection, and classified by a random forest. Extensive experiments on challenging datasets show that the proposed algorithm performs favorably against the state-of-the-art methods on head pose estimation in low-resolution images degraded by noise, occlusion, and blurring.

****

An MRF-Poselets Model for Detecting Highly Articulated Humans

Duc Thanh Nguyen, Minh-Khoi Tran, Sai-Kit Yeung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1967-1975

Detecting highly articulated objects such as humans is a challenging problem. This paper proposes a novel part-based model built upon poselets, a notion of parts, and Markov Random Field (MRF) for modelling the human body structure under the variation of human poses and viewpoints. The problem of human detection is then formulated as maximum a posteriori (MAP) estimation in the MRF model. Variational mean field method, a robust statistical inference, is adopted to approximate the MAP estimation. The proposed method was evaluated and compared with existing methods on different test sets including H3D and PASCAL VOC 2007-2009. Experimental results have favourbly shown the robustness of the proposed method in comparison to the state-of-the-art.

****

Beyond Tree Structure Models: A New Occlusion Aware Graphical Model for Human Pose Estimation

Lianrui Fu, Junge Zhang, Kaiqi Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1976-1984

Occlusion is a main challenge for human pose estimation, which is largely ignored in popular tree structure models. The tree structure model is simple and convenient for exact inference, but short in modeling the occlusion coherence especially in the case of self-occlusion. We propose an occlusion aware graphical model which is able to model both self-occlusion and occlusion by the other objects simultaneously. The proposed model structure can encode the interactions between human body parts and objects, and hence enable it to learn occlusion coherence from data discriminatively. We evaluate our model on several public benchmarks for human pose estimation including challenging subsets featuring significant occlusion. The experimental results show that our method obtains comparable accuracy with the state-of-the-arts, and achieves promising performance in 2D human pose estimation with occlusion.

****

Relaxing From Vocabulary: Robust Weakly-Supervised Deep Learning for Vocabulary-Free Image Tagging

Jianlong Fu, Yue Wu, Tao Mei, Jinqiao Wang, Hanqing Lu, Yong Rui; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1985-1993

The development of deep learning has empowered machines with comparable capability of recognizing limited image categories to human beings. However, most existing approaches heavily rely on human-curated training data, which hinders the scalability to large and unlabeled vocabularies in image tagging. In this paper, we propose a weakly-supervised deep learning model which can be trained from the readily available Web images to relax the dependence on human labors and scale up to arbitrary tags (categories). Specifically, based on the assumption that features of true samples in a category tend to be similar and noises tend to be variant, we embed the feature map of the last deep layer into a new affinity representation, and further minimize the discrepancy between the affinity representation and its low-rank approximation. The discrepancy is finally transformed into the objective function to give relevance feedback to back propagation. Experiments show that we can achieve a performance gain of 14.0% in terms of a semantic-based relevance metric in image tagging with 63,043 tags from the WordNet, against the typical deep model trained on the ImageNet 1,000 vocabulary set.

****

Visual Phrases for Exemplar Face Detection

Vijay Kumar, Anoop Namboodiri, C. V. Jawahar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1994-2002

Recently, exemplar based approaches have been successfully applied for face detection in the wild. Contrary to traditional approaches that model face variations from a large and diverse set of training examples, exemplar-based approaches use a collection of discriminatively trained exemplars for detection. In this para digm, each exemplar casts a vote using retrieval framework and generalized Hough voting, to locate the faces in the target image. The advantage of this approach is that by having a large database that covers all possible variations, faces in challenging conditions can be detected without having to learn explicit models for different variations. Current schemes, however, make an assumption of independence between the visual words, ignoring their relations in the process. They also ignore the spatial consistency of the visual words. Consequently, every exemplar word contributes equally during voting regardless of its location. In this paper, we propose a novel approach that incorporates higher order information in the voting process. We discover visual phrases that contain semantically related visual words and exploit them for detection along with the visual words. For spatial consistency, we estimate the spatial distribution of visual words and phrases from the entire database and then weigh their occurrence in exemplars. This ensures that a visual word or a phrase in an exemplar makes a major contribution only if it occurs at its semantic location, thereby suppressing the noise significantly. We perform extensive experiments on standard FDDB, AFW and G-album datasets and show significant improvement over previous exemplar approaches.

****

Spatial Semantic Regularisation for Large Scale Object Detection
Damian Mrowca, Marcus Rohrbach, Judy Hoffman, Ronghang Hu, Kate Saenko, Trevor Darrell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2003-2011

Large scale object detection with thousands of classes introduces the problem of many contradicting false positive detections, which have to be suppressed. Class-independent non-maximum suppression has traditionally been used for this step, but it does not scale well as the number of classes grows. Traditional non-maximum suppression does not consider label- and instance-level relationships nor does it allow an exploitation of the spatial layout of detection proposals. We propose a new multi-class spatial semantic regularisation method based on affinity propagation clustering, which simultaneously optimises across all categories and all proposed locations in the image, to improve both the localisation and categorisation of selected detection proposals. Constraints are shared across the labels through the semantic WordNet hierarchy. Our approach proves to be especially useful in large scale settings with thousands of classes, where spatial and semantic interactions are very frequent and only weakly supervised detectors can be built due to a lack of bounding box annotations. Detection experiments are conducted on the ImageNet and COCO dataset, and in settings with thousands of detected categories. Our method provides a significant precision improvement by reducing false positives, while simultaneously improving the recall.

****

Human Pose Estimation in Videos
Dong Zhang, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2012-2020

In this paper, we present a method to estimate a sequence of human poses in unconstrained videos. In contrast to the commonly employed graph optimization framework, which is NP-hard and needs approximate solutions, we formulate this problem into a unified two stage tree-based optimization problem for which an efficient and exact solution exists. Although the proposed method finds an exact solution, it does not sacrifice the ability to model the spatial and temporal constraints between body parts in the video frames; indeed it even models the symmetric parts better than the existing methods. The proposed method is based on two main ideas: `Abstraction' and `Association' to enforce the intra- and inter-frame body part constraints respectively without inducing extra computational complexity to the polynomial time solution. Using the idea of `Abstraction', a new concept of `abstract body part' is introduced to model not only the tree based body part structure similar to existing methods, but also extra constraints between symmetric parts. Using the idea of `Association', the optimal tracklets are generated for each abstract body part, in order to enforce the spatiotemporal constraints between body parts in adjacent frames. Finally, a sequence of the best poses is inferred from the abstract body part tracklets through the tree-based optimizati

on. We evaluated the proposed method on three publicly available video based human pose estimation datasets, and obtained dramatically improved performance compared to the state-of-the-art methods.
****

Contour Box: Rejecting Object Proposals Without Explicit Closed Contours
Cewu Lu, Shu Liu, Jiaya Jia, Chi-Keung Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2021-2029
Closed contour is an important objectness indicator. We propose a new measure subject to the completeness and tightness constraints, where the optimized closed contour should be tightly bounded within an object proposal. The closed contour measure is defined using closed path integral, and we solve the optimization problem efficiently in polar coordinate system with a global optimum guaranteed. Extensive experiments show that our method can reject a large number of false proposals, and achieve over 6% improvement in object recall at the challenging overlap threshold 0.8 on the VOC 2007 test dataset.
****

Registering Images to Untextured Geometry Using Average Shading Gradients
Tobias Plotz, Stefan Roth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2030-2038
Many existing approaches for image-to-geometry registration assume that either a textured 3D model or a good initial guess of the 3D pose is available to bootstrap the registration process. In this paper we consider the registration of photographs to 3D models even when no texture information is available. This is very challenging as we cannot rely on texture gradients, and even shading gradients are hard to estimate since the lighting conditions are unknown. To that end, we propose average shading gradients, a rendering technique that estimates the average gradient magnitude over all lighting directions under Lambertian shading. We use this gradient representation as the building block of a registration pipeline based on matching sparse features. To cope with inevitable false matches due to the missing texture information and to increase robustness, the pose of the 3D model is estimated in two stages. Coarse pose hypotheses are first obtained from a single correct match each, subsequently refined using SIFT flow, and finally verified. We apply our algorithm to registering images of real-world objects to untextured 3D meshes of limited accuracy.
****

Robust Nonrigid Registration by Convex Optimization
Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2039-2047
We present an approach to nonrigid registration of 3D surfaces. We cast isometric embedding as MRF optimization and apply efficient global optimization algorithms based on linear programming relaxations. The Markov random field perspective suggests a natural connection with robust statistics and motivates robust forms of the intrinsic distortion functional. Our approach outperforms a large body of prior work by a significant margin, increasing registration precision on real data by a factor of 3.
****

Robust and Optimal Sum-of-Squares-Based Point-to-Plane Registration of Image Sets and Structured Scenes
Danda Pani Paudel, Adlane Habed, Cedric Demonceaux, Pascal Vasseur; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2048-2056
This paper deals with the problem of registering a known structured 3D scene and its metric Structure-from-Motion (SfM) counterpart. The proposed work relies on a prior plane segmentation of the 3D scene and aligns the data obtained from both modalities by solving the point-to-plane assignment problem. An inliers-maximization approach within a Branch-and-Bound (BnB) search scheme is adopted. For the first time in this paper, a Sum-of-Squares optimization theory framework is employed for identifying point-to-plane mismatches (i.e. outliers) with certainty . This allows us to iteratively build potential inliers sets and converge to the solution satisfied by the largest number of point-to-plane assignments. Further

more, our approach is boosted by new plane visibility conditions which are also introduced in this paper. Using this framework, we solve the registration problem in two cases: (i) a set of putative point-to-plane correspondences (with possibly overwhelmingly many outliers) is given as input and (ii) no initial correspondences are given. In both cases, our approach yields outstanding results in terms of robustness and optimality.

****

MeshStereo: A Global Stereo Model With Mesh Alignment Regularization for View Interpolation

Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, Yong Rui; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2057-2065

We present a novel global stereo model designed for view interpolation. Unlike existing stereo models which only output a disparity map, our model is able to output a 3D triangular mesh, which can be directly used for view interpolation. To this aim, we partition the input stereo images into 2D triangles with shared vertices. Lifting the 2D triangulation to 3D naturally generates a corresponding mesh. A technical difficulty is to properly split vertices to multiple copies when they appear at depth discontinuous boundaries. To deal with this problem, we formulate our objective as a two-layer MRF, with the upper layer modeling the splitting properties of the vertices and the lower layer optimizing a region-based stereo matching. Experiments on the Middlebury and the Herodion datasets demonstrate that our model is able to synthesize visually coherent new view angles with high PSNR, as well as outputting high quality disparity maps which rank at the first place on the new challenging high resolution Middlebury 3.0 benchmark.

****

CV-HAZOP: Introducing Test Data Validation for Computer Vision

Oliver Zendel, Markus Murschitz, Martin Humenberger, Wolfgang Herzner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2066-2074

Test data plays an important role in computer vision (CV) but is plagued by two questions: Which situations should be covered by the test data and have we tested enough to reach a conclusion? In this paper we propose a new solution answering these questions using a standard procedure devised by the safety community to validate complex systems: The Hazard and Operability Analysis (HAZOP). It is designed to systematically search and identify difficult, performance-decreasing situations and aspects. We introduce a generic CV model that creates the basis for the hazard analysis and, for the first time, apply an extensive HAZOP to the CV domain. The result is a publicly available checklist with more than 900 identified individual hazards. This checklist can be used to evaluate existing test datasets by quantifying the amount of covered hazards. We evaluate our approach by first analyzing and annotating the popular stereo vision test datasets Middlebury and KITTI. Second, we compare the performance of six popular stereo matching algorithms at the identified hazards from our checklist with their average performance and show, as expected, a clear negative influence of the hazards. The presented approach is a useful tool to evaluate and improve test datasets and creates a common basis for future dataset designs.

****

Structure From Motion Using Structure-Less Resection

Enliang Zheng, Changchang Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2075-2083

This paper proposes a new incremental structure from motion (SfM) algorithm based on a novel structure-less camera resection technique. Traditional methods rely on 2D-3D correspondences to compute the pose of candidate cameras using PnP. In this work, we take the collection of already reconstructed cameras as a generalized camera, and determine the absolute pose of a candidate pinhole camera from pure 2D correspondences, which we call it semi-generalized camera pose problem. We present the minimal solvers of the new problem for both calibrated and partially calibrated (unknown focal length) pinhole cameras. By integrating these new algorithms in an incremental SfM system, we go beyond the state-of-art method s with the capability of reconstructing cameras without 2D-3D correspondences. Large-scale real image experiments show that our new SfM system significantly improves the completeness of 3D reconstruction over the standard approach.

****

Joint Camera Clustering and Surface Segmentation for Large-Scale Multi-View Stereo

Runze Zhang, Shiwei Li, Tian Fang, Siyu Zhu, Long Quan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2084-2092

In this paper, we propose an optimal decomposition approach to large-scale multi-view stereo from an initial sparse reconstruction. The success of the approach depends on the introduction of surface-segmentation-based camera clustering rather than sparse-point-based camera clustering, which suffers from the problems of non-uniform reconstruction coverage ratio and high redundancy. In details, we introduce three criteria for camera clustering and surface segmentation for reconstruction, and then we formulate these criteria into an energy minimization problem under constraints. To solve this problem, we propose a joint optimization in a hierarchical framework to obtain the final surface segments and corresponding optimal camera clusters. On each level of the hierarchical framework, the camera clustering problem is formulated as a parameter estimation problem of a probability model solved by a General Expectation-Maximization algorithm and the surface segmentation problem is formulated as a Markov Random Field model based on the probability estimated by the previous camera clustering process. The experiments on several Internet datasets and aerial photo datasets demonstrate that the proposed approach method generates more uniform and complete dense reconstruction with less redundancy, resulting in more efficient multi-view stereo algorithm.

****

Higher-Order CRF Structural Segmentation of 3D Reconstructed Surfaces

Jingbo Liu, Jinglu Wang, Tian Fang, Chiew-Lan Tai, Long Quan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2093-2101

In this paper, we propose a structural segmentation algorithm to partition multi-view stereo reconstructed surfaces of large-scale urban environments into structural segments. Each segment corresponds to a structural component describable by a surface primitive of up to the second order. This segmentation is for use in subsequent urban object modeling, vectorization, and recognition. To overcome the high geometrical and topological noise levels in the 3D reconstructed urban surfaces, we formulate the structural segmentation as a higher-order Conditional Random Field (CRF) labeling problem. It not only incorporates classical lower-order 2D and 3D local cues, but also encodes contextual geometric regularities to disambiguate the noisy local cues. A general higher-order CRF is difficult to solve. We develop a bottom-up progressive approach through a patch-based surface representation, which iteratively evolves from the initial mesh triangles to the final segmentation. Each iteration alternates between performing a prior discovery step, which finds the contextual regularities of the patch-based representation, and an inference step that leverages the regularities as higher-order priors to construct a more stable and regular segmentation. The efficiency and robustness of the proposed method is extensively demonstrated on real reconstruction models, yielding significantly better performance than classical mesh segmentation methods.

****

Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition

Torsten Sattler, Michal Havlena, Filip Radenovic, Konrad Schindler, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2102-2110

Structure-based localization is the task of finding the absolute pose of a given query image w.r.t. a pre-computed 3D model. While this is almost trivial at small scale, special care must be taken as the size of the 3D model grows, because straight-forward descriptor matching becomes ineffective due to the large memory footprint of the model, as well as the strictness of the ratio test in 3D. Recently, several authors have tried to overcome these problems, either by a smart compression of the 3D model or by clever sampling strategies for geometric verifi cation. Here we explore an orthogonal strategy, which uses all the 3D points and standard sampling, but performs feature matching implicitly, by quantization into a fine vocabulary. We show that although this matching is ambiguous and gives rise to 3D hyperpoints when matching each 2D query feature in isolation, a simple voting strategy, which enforces the fact that the selected 3D points shall be co-visible, can reliably find a locally unique 2D-3D point assignment. Experiments on two large-scale datasets demonstrate that our method achieves state-of-the-art performance, while the memory footprint is greatly reduced, since only visual word labels but no 3D point descriptors need to be stored.

****

Globally Optimal 2D-3D Registration From Points or Lines Without Correspondences
Mark Brown, David Windridge, Jean-Yves Guillemaut; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2111-2119

We present a novel approach to 2D-3D registration from points or lines without correspondences. While there exist established solutions in the case where correspondences are known, there are many situations where it is not possible to reliably extract such correspondences across modalities, thus requiring the use of a correspondence-free registration algorithm. Existing correspondence-free methods rely on local search strategies and consequently have no guarantee of finding the optimal solution. In contrast, we present the first globally optimal approach to 2D-3D registration without correspondences, achieved by a Branch-and-Bound algorithm. Furthermore, a deterministic annealing procedure is proposed to speed up the nested branch-and-bound algorithm used. The theoretical and practical advantages this brings are demonstrated on a range of synthetic and real data where it is observed that the proposed approach is significantly more robust to high proportions of outliers compared to existing approaches.

****

The HCI Stereo Metrics: Geometry-Aware Performance Analysis of Stereo Algorithms
Katrin Honauer, Lena Maier-Hein, Daniel Kondermann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2120-2128

Performance characterization of stereo methods is mandatory to decide which algorithm is useful for which application. Prevalent benchmarks mainly use the root mean squared error (RMS) with respect to ground truth disparity maps to quantify algorithm performance. We show that the RMS is of limited expressiveness for algorithm selection and introduce the HCI Stereo Metrics. These metrics assess stereo results by harnessing three semantic cues: depth discontinuities, planar surfaces, and fine geometric structures. For each cue, we extract the relevant set of pixels from existing ground truth. We then apply our evaluation functions to quantify characteristics such as edge fattening and surface smoothness. We demonstrate that our approach supports practitioners in selecting the most suitable algorithm for their application. Using the new Middlebury dataset, we show that rankings based on our metrics reveal specific algorithm strengths and weaknesses which are not quantified by existing metrics. We finally show how stacked bar charts and radar charts visually support multidimensional performance evaluation. An interactive stereo benchmark based on the proposed metrics and visualizations is available at: http://hci.iwr.uni-heidelberg.de/stereometrics

****

Merging the Unmatchable: Stitching Visually Disconnected SfM Models
Andrea Cohen, Torsten Sattler, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2129-2137

Recent advances in Structure-from-Motion not only enable the reconstruction of large scale scenes, but are also able to detect ambiguous structures caused by repeating elements that might result in incorrect reconstructions. Yet, it is not always possible to fully reconstruct a scene. The images required to merge different sub-models might be missing or it might be impossible to acquire such images in the first place due to occlusions or the structure of the scene. The problem of aligning multiple reconstructions that do not have visual overlap is impossible to solve in general. An important variant of this problem is the case in which individual sides of a building can be reconstructed but not joined due to the missing visual overlap. In this paper, we present a combinatorial approach for solving this variant by automatically stitching multiple sides of a building together. Our approach exploits symmetries and semantic information to reason about the possible geometric relations between the individual models. We show that our approach is able to reconstruct complete building models where traditional SfM ends up with disconnected building sides.

****

3D Fragment Reassembly Using Integrated Template Guidance and Fracture-Region Matching

Kang Zhang, Wuyi Yu, Mary Manhein, Warren Waggenspack, Xin Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2138-2146

This paper studies matching of fragmented objects to recompose their original geometry. Solving this geometric reassembly problem has direct applications in archaeology and forensic investigation in the computer-aided restoration of damaged artifacts and evidence. We develop a new algorithm to effectively integrate both guidance from a template and from matching of adjacent pieces' fracture-regions. First, we compute partial matchings between fragments and a template, and pairwise matchings among fragments. Many potential matches are obtained and then selected/refined in a multi-piece matching stage to maximize global groupwise matching consistency. This pipeline is effective in composing fragmented thin-shell objects containing small pieces, whose pairwise matching is usually unreliable and ambiguous and hence their reassembly remains challenging to the existing algorithms.

****

Procedural Editing of 3D Building Point Clouds

Ilke Demir, Daniel G. Aliaga, Bedrich Benes; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2147-2155

Thanks to the recent advances in computational photography and remote sensing, point clouds of buildings are becoming increasingly available, yet their processing poses various challenges. In our work, we tackle the problem of point cloud completion and editing and we approach it via inverse procedural modeling. Contrary to the previous work, our approach operates directly on the point cloud without an intermediate triangulation. Our approach consists of 1) semi-automatic segmentation of the input point cloud with segment comparison and template matching to detect repeating structures, 2) a consensus-based voting schema and a pattern extraction algorithm to discover completed terminal geometry and their patterns of usage, all encoded into a context-free grammar, and 3) an interactive editing tool where the user can create new point clouds by using procedural copy and paste operations, and smart resizing. We demonstrate our approach on editing of building models with up to 1.8M points. In our implementation, preprocessing takes up to several minutes and a single editing operation needs from one second to one minute depending on the model size and the operation type.

****

Semantically-Aware Aerial Reconstruction From Multi-Modal Data

Randi Cabezas, Julian Straub, John W. Fisher III; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2156-2164

We consider a methodology for integrating multiple sensors along with semantic information to enhance scene representations. We propose a probabilistic generative model for inferring semantically-informed aerial reconstructions from multi-modal data within a consistent mathematical framework. The approach, called Semantically- Aware Aerial Reconstruction (SAAR), not only exploits inferred scene geometry, appearance, and semantic observations to obtain a meaningful categorization of the data, but also extends previously proposed methods by imposing structure on the prior over geometry, appearance, and semantic labels. This leads to more accurate reconstructions and the ability to fill in missing contextual labels via joint sensor and semantic information. We introduce a new multi-modal synthetic dataset in order to provide quantitative performance analysis. Additionally, we apply the model to real-world data and exploit OpenStreetMap as a source of semantic observations. We show quantitative improvements in reconstruction accuracy of large-scale urban scenes from the combination of LiDAR, aerial photography, and semantic data. Furthermore, we demonstrate the model's ability to f

ill in for missing sensed data, leading to more interpretable reconstructions.
****

Guaranteed Outlier Removal for Rotation Search
Alvaro Parra Bustos, Tat-Jun Chin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2165-2173

Rotation search has become a core routine for solving many computer vision problems. The aim is to rotationally align two input point sets with correspondences. Recently, there is significant interest in developing globally optimal rotation search algorithms. A notable weakness of global algorithms, however, is their relatively high computational cost, especially on large problem sizes and data with a high proportion of outliers. In this paper, we propose a novel outlier removal technique for rotation search. Our method guarantees that any correspondence it discards as an outlier does not exist in the inlier set of the globally optimal rotation for the original data. Based on simple geometric operations, our algorithm is deterministic and fast. Used as a preprocessor to prune a large portion of the outliers from the input data, our method enables substantial speed-up of rotation search algorithms without compromising global optimality. We demonstrate the efficacy of our method in various synthetic and real data experiments.
****

Peeking Template Matching for Depth Extension
Simon Korman, Eyal Ofek, Shai Avidan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2174-2182

We propose a method that extends a given depth image into regions in 3D that are not visible from the point of view of the camera. The algorithm detects repeated 3D structures in the visible scene and suggests a set of 3D extension hypotheses, which are then combined together through a global 3D MRF discrete optimization. The recovered global 3D surface is consistent with both the input depth map and the hypotheses. A key component of this work is a novel 3D template matcher that is used to detect repeated 3D structure in the scene and to suggest the hypotheses. A unique property of this matcher is that it can handle depth uncertainty. This is crucial because the matcher is required to ``peek around the corner'', as it operates at the boundaries of the visible 3D scene where depth information is missing. The proposed matcher is fast and is guaranteed to find an approximation to the globally optimal solution. We demonstrate on real-world data that our algorithm is capable of completing a full 3D scene from a single depth image and can synthesize a full depth map from a novel viewpoint of the scene. In addition, we report results on an extensive synthetic set of 3D shapes, which allows us to evaluate the method both qualitatively and quantitatively.
****

Deformable 3D Fusion: From Partial Dynamic 3D Observations to Complete 4D Models
Weipeng Xu, Mathieu Salzmann, Yongtian Wang, Yue Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2183-2191

Capturing the 3D motion of dynamic, non-rigid objects has attracted significant attention in computer vision. Existing methods typically require either complete 3D volumetric observations, or a shape template. In this paper, we introduce a template-less 4D reconstruction method that incrementally fuses highly-incomplete 3D observations of a deforming object, and generates a complete, temporally-coherent shape representation of the object. To this end, we design an online algorithm that alternatively registers new observations to the current model estimate and updates the model. We demonstrate the effectiveness of our approach at reconstructing non-rigidly moving objects from highly-incomplete measurements on both sequences of partial 3D point clouds and Kinect videos.
****

Non-Parametric Structure-Based Calibration of Radially Symmetric Cameras
Federico Camposeco, Torsten Sattler, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2192-2200

We propose a novel two-step method for estimating the intrinsic and extrinsic calibration of any radially symmetric camera, including non-central systems. The first step consists of estimating the camera pose, given a Structure from Motion (SfM) model, up to the translation along the optical axis. As a second step, we obtain the calibration by finding the translation of the camera center using an ordering constraint. The method makes use of the 1D radial camera model, which allows us to effectively handle any radially symmetric camera, including non-central ones. Using this ordering constraint, we show that the we are able to calibrate several different (central and non-central) Wide Field of View (WFOV) cameras, including fisheye, hyper-catadioptric and spherical catadioptric cameras, as well as pinhole cameras, using a single image or jointly solving for several views.

****

Exploiting Object Similarity in 3D Reconstruction

Chen Zhou, Fatma Guney, Yizhou Wang, Andreas Geiger; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2201-2209

Despite recent progress, reconstructing outdoor scenes in 3D from movable platforms remains a highly difficult endeavour. Challenges include low frame rates, occlusions, large distortions and difficult lighting conditions. In this paper, we leverage the fact that the larger the reconstructed area, the more likely objects of similar type and shape will occur in the scene. This is particularly true for outdoor scenes where buildings and vehicles often suffer from missing texture or reflections, but share similarity in 3D shape. We take advantage of this shape similarity by localizing objects using detectors and jointly reconstructing them while learning a volumetric model of their shape. This allows us to reduce noise while completing missing surfaces as objects of similar shape benefit from all observations for the respective category. We evaluate our approach with respect to LIDAR ground truth on a novel challenging suburban dataset and show its advantages over the state-of-the-art.

****

You Are Here: Mimicking the Human Thinking Process in Reading Floor-Plans

Hang Chu, Dong Ki Kim, Tsuhan Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2210-2218

A human can easily find his or her way in an unfamiliar building, by walking around and reading the floor-plan. We try to mimic and automate this human thinking process. More precisely, we introduce a new and useful task of locating an user in the floor-plan, by using only a camera and a floor-plan without any other prior information. We address the problem with a novel matching-localization algorithm that is inspired by human logic. We demonstrate through experiments that our method outperforms state-of-the-art floor-plan-based localization methods by a large margin, while also being highly efficient for real-time applications.

****

MAP Disparity Estimation Using Hidden Markov Trees

Eric T. Psota, Jedrzej Kowalczuk, Mateusz Mittek, Lance C. Perez; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2219-2227

A new method is introduced for stereo matching that operates on minimum spanning trees (MSTs) generated from the images. Disparity maps are represented as a collection of hidden states on MSTs, and each MST is modeled as a hidden Markov tree. An efficient recursive message-passing scheme designed to operate on hidden Markov trees, known as the upward-downward algorithm, is used to compute the maximum a posteriori (MAP) disparity estimate at each pixel. The messages processed by the upward-downward algorithm involve two types of probabilities: the probability of a pixel having a particular disparity given a set of per-pixel matching costs, and the probability of a disparity transition between a pair of connected pixels given their similarity. The distributions of these probabilities are modeled from a collection of images with ground truth disparities. Performance evaluation using the Middlebury stereo benchmark version 3 demonstrates that the proposed method ranks second and third in terms of overall accuracy when evaluated on the training and test image sets, respectively.

****

Wide Baseline Stereo Matching With Convex Bounded Distortion Constraints

Meirav Galun, Tal Amir, Tal Hassner, Ronen Basri, Yaron Lipman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2228-2236

Finding correspondences in wide baseline setups is a challenging problem. Existing approaches have focused largely on developing better feature descriptors for correspondence and on accurate recovery of epipolar line constraints. This paper focuses on the challenging problem of finding correspondences once approximate epipolar constraints are given. We introduce a novel method that integrates a deformation model. Specifically, we formulate the problem as finding the largest number of corresponding points related by a bounded distortion map that obeys the given epipolar constraints. We show that, while the set of bounded distortion maps is not convex, the subset of maps that obey the epipolar line constraints is convex, allowing us to introduce an efficient algorithm for matching. We further utilize a robust cost function for matching and employ majorization-minimization for its optimization. Our experiments indicate that our method finds significantly more accurate maps than existing approaches.
****

Interactive Visual Hull Refinement for Specular and Transparent Object Surface Reconstruction

Xinxin Zuo, Chao Du, Sen Wang, Jiangbin Zheng, Ruigang Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2237-2245

In this paper we present a method of using standard multi-view images for 3D surface reconstruction of non-Lambertian objects. We extend the original visual hull concept to incorporate 3D cues presented by internal occluding contours, i.e., occluding contours that are inside the object's silhouettes. We discovered that these internal contours, which are results of convex parts on an object's surface, can lead to a tighter fit than the original visual hull. We formulated a new visual hull refinement scheme - Locally Convex Carving that can completely reconstruct concavity caused by two or more intersecting convex surfaces. In addition we develop a novel approach for contour tracking given labeled contours in sparse key frames. It is designed specifically for highly specular or transparent objects, for which assumptions made in traditional contour detection/tracking methods, such as highest gradient and stationary texture edges, are no longer valid. It is formulated as an energy minimization function where several novel terms are developed to increase robustness. Based on the two core algorithms, we have developed an interactive system for 3D modeling. We have validated our system, both quantitatively and qualitatively, with four datasets of different object materials. Results show that we are able to generate visually pleasing models for very challenging cases.
****

Hierarchical Higher-Order Regression Forest Fields: An Application to 3D Indoor Scene Labelling

Trung T. Pham, Ian Reid, Yasir Latif, Stephen Gould; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2246-2254

This paper addresses the problem of semantic segmentation of 3D indoor scenes reconstructed from RGB-D images.Traditionally label prediction for 3D points is tackled by employing graphical models that capture scene features and complex relations between different class labels. However, the existing work is restricted to pairwise conditional random fields, which are insufficient when encoding rich scene context. In this work we propose models with higher-order potentials to describe complex relational information from the 3D scenes. Specifically, we relax the labelling problem to a regression, and generalize the higher-order associative P n Potts model to a new family of arbitrary higher-order models based on regression forests. We show that these models, like the robust P n models, can still be decomposed into the sum of pairwise terms by introducing auxiliary variables. Moreover, our proposed higher-order models also permit extension to hierarchical random fields, which allows for the integration of scene context and features computed at different scales. Our potential functions are constructed based on regression forests encoding Gaussian densities that admit efficient inference. The parameters of our model are learned from training data using a structured learning approach. Results on two datasets show clear improvements over current state-of-the-art methods.
****

Classical Scaling Revisited
Gil Shamai, Yonathan Aflalo, Michael Zibulevsky, Ron Kimmel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2255-2263
Multidimensional-scaling (MDS) is an information analysis tool. It involves the evaluation of distances between data points, which is a quadratic space-time problem. Then, MDS procedures find an embedding of the points in a low dimensional Euclidean (flat) domain, optimizing for the similarity of inter-points distances. We present an efficient solver for Classical Scaling (a specific MDS model) by extending the distances measured from a subset of the points to the rest, while exploiting the smoothness property of the distance functions. The smoothness is measured by the L2 norm of the Laplace-Beltrami operator applied to the unknown distance function. The Laplace Beltrami reflects the local differential relations between points, and can be computed in linear time. Classical-scaling is thereby reformulated into a quasi-linear space-time complexities procedure.
****
Dense Continuous-Time Tracking and Mapping With Rolling Shutter RGB-D Cameras
Christian Kerl, Jorg Stuckler, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2264-2272
We propose a dense continuous-time tracking and mapping method for RGB-D cameras. We parametrize the camera trajectory using continuous B-splines and optimize the trajectory through dense, direct image alignment. Our method also directly models rolling shutter in both RGB and depth images within the optimization, which improves tracking and reconstruction quality for low-cost CMOS sensors. Using a continuous trajectory representation has a number of advantages over a discrete-time representation (e.g. camera poses at the frame interval). With splines, less variables need to be optimized than with a discrete representation, since the trajectory can be represented with fewer control points than frames. Splines also naturally include smoothness constraints on derivatives of the trajectory estimate. Finally, the continuous trajectory representation allows to compensate for rolling shutter effects, since a pose estimate is available at any exposure time of an image. Our approach demonstrates superior quality in tracking and reconstruction compared to approaches with discrete-time or global shutter assumptions.
****
Dense Image Registration and Deformable Surface Reconstruction in Presence of Occlusions and Minimal Texture
Dat Tien Ngo, Sanghyuk Park, Anne Jorstad, Alberto Crivellaro, Chang D. Yoo, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2273-2281
Deformable surface tracking from monocular images is well-known to be under-constrained. Occlusions often make the task even more challenging, and can result in failure if the surface is not sufficiently textured. In this work, we explicitly address the problem of 3D reconstruction of poorly textured, occluded surfaces, proposing a framework based on a template-matching approach that scales dense robust features by a relevancy score. Our approach is extensively compared to current methods employing both local feature matching and dense template alignment. We test on standard datasets as well as on a new dataset (that will be made publicly available) of a sparsely textured, occluded surface. Our framework achieves state-of-the-art results for both well and poorly textured, occluded surfaces.
****
The Likelihood-Ratio Test and Efficient Robust Estimation
Andrea Cohen, Christopher Zach; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2282-2290
Robust estimation of model parameters in the presence of outliers is a key problem in computer vision. RANSAC inspired techniques are widely used in this context, although their application might be limited due to the need of a priori knowledge on the inlier noise level. We propose a new approach for jointly optimizing over model parameters and the inlier noise level based on the likelihood ratio test. This allows control over the type I error incurred. We also propose an ear

ly bailout strategy for efficiency. Tests on both synthetic and real data show that our method outperforms the state-of-the-art in a fraction of the time.

****

Reflection Modeling for Passive Stereo

Rahul Nair, Andrew Fitzgibbon, Daniel Kondermann, Carsten Rother; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2291-2299

Stereo reconstruction in presence of reality faces many challenges that still need to be addressed. This paper considers reflections, which introduce incorrect matches due to the observation violating the diffuse-world assumption underlying the majority of stereo techniques. Unlike most existing work, which employ regularization or robust data terms to suppress such errors, we derive two least squares models from first principles that generalize diffuse world stereo and explicitly take reflections into account.   These models are parametrized by depth, orientation and material properties, resulting in a total of up to 5 parameters per pixel that have to be estimated. Additionally large non-local interactions between viewed and reflected surface have to be taken into account. These two properties make inference of the model appear prohibitive, but we present evidence that inference is actually possible using a variant of patch match stereo.

****

Detailed Full-Body Reconstructions of Moving People From Monocular RGB-D Sequences

Federica Bogo, Michael J. Black, Matthew Loper, Javier Romero; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2300-2308

We accurately estimate the 3D geometry and appearance of the human body from a monocular RGB-D sequence of a user moving freely in front of the sensor. Range data in each frame is first brought into alignment with a multi-resolution 3D body model in a coarse-to-fine process. The method then uses geometry and image texture over time to obtain accurate shape, pose, and appearance information despite unconstrained motion, partial views, varying resolution, occlusion, and soft tissue deformation. Our novel body model has variable shape detail, allowing it to capture faces with a high-resolution deformable head model and body shape with lower-resolution. Finally we combine range data from an entire sequence to estimate a high-resolution displacement map that captures fine shape details. We compare our recovered models with high-resolution scans from a professional system and with avatars created by a commercial product. We extract accurate 3D avatars from challenging motion sequences and even capture soft tissue dynamics.

****

Efficient Solution to the Epipolar Geometry for Radially Distorted Cameras

Zuzana Kukelova, Jan Heller, Martin Bujnak, Andrew Fitzgibbon, Tomas Pajdla; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2309-2317

The estimation of the epipolar geometry of two cameras from image matches is a fundamental problem of computer vision with many applications. While the closely related problem of estimating relative pose of two different uncalibrated cameras with radial distortion is of particular importance, none of the previously published methods is suitable for practical applications. These solutions are either numerically unstable, sensitive to noise, based on a large number of point correspondences, or simply too slow for real-time applications. In this paper, we present a new efficient solution to this problem that uses 10 image correspondences. By manipulating ten input polynomial equations, we derive a degree 10 polynomial equation in one variable. The solutions to this equation are efficiently found using the Sturm sequences method. In the experiments, we show that the proposed solution is stable, noise resistant, and fast, and as such efficiently usable in a practical Structure-from-Motion pipeline.

****

Learning a Descriptor-Specific 3D Keypoint Detector

Samuele Salti, Federico Tombari, Riccardo Spezialetti, Luigi Di Stefano; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2318-2326

Keypoint detection represents the first stage in the majority of modern computer vision pipelines based on automatically established correspondences between local descriptors. However, no standard solution has emerged yet in the case of 3D data such as point clouds or meshes, which exhibit high variability in level of detail and noise. More importantly, existing proposals for 3D keypoint detection rely on geometric saliency functions that attempt to maximize repeatability rather than distinctiveness of the selected regions, which may lead to sub-optimal performance of the overall pipeline. To overcome these shortcomings, we cast 3D keypoint detection as a binary classification between points whose support can be correctly matched by a predefined 3D descriptor or not, thereby learning a descriptor-specific detector that adapts seamlessly to different scenarios. Through experiments on several public datasets, we show that this novel approach to the design of a keypoint detector represents a flexible solution that, nonetheless, can provide state-of-the-art descriptor matching performance.

****

Component-Wise Modeling of Articulated Objects

Valsamis Ntouskos, Marta Sanzari, Bruno Cafaro, Federico Nardi, Fabrizio Natola, Fiora Pirri, Manuel Ruiz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2327-2335

We introduce a novel framework for modeling articulated objects based on the aspects of their components. By decomposing the object into components, we divide the problem in smaller modeling tasks. After obtaining 3D models for each component aspect by employing a shape deformation paradigm, we merge them together, forming the object components. The final model is obtained by assembling the components using an optimization scheme which fits the respective 3D models to the corresponding apparent contours in a reference pose. The results suggest that our approach can produce realistic 3D models of articulated objects in reasonable time.

****

A Collaborative Filtering Approach to Real-Time Hand Pose Estimation

Chiho Choi, Ayan Sinha, Joon Hee Choi, Sujin Jang, Karthik Ramani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2336-2344

Collaborative filtering aims to predict unknown user ratings in a recommender system by collectively assessing known user preferences. In this paper, we first draw analogies between collaborative filtering and the pose estimation problem. Specifically, we recast the hand pose estimation problem as the cold-start problem for a new user with unknown item ratings in a recommender system. Inspired by fast and accurate matrix factorization techniques for collaborative filtering, we develop a real-time algorithm for estimating the hand pose from RGB-D data of a commercial depth camera. First, we efficiently identify nearest neighbors using local shape descriptors in the RGB-D domain from a library of hand poses with known pose parameter values. We then use this information to evaluate the unknown pose parameters using a joint matrix factorization and completion (JMFC) approach. Our quantitative and qualitative results suggest that our approach is robust to variation in hand configurations while achieving real time performance (29 FPS) on a standard computer.

****

On the Equivalence of Moving Entrance Pupil and Radial Distortion for Camera Calibration

Avinash Kumar, Narendra Ahuja; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2345-2353

Radial distortion for ordinary (non-fisheye) camera lenses has traditionally been modeled as an infinite series function of radial location of an image pixel from the image center. While there has been enough empirical evidence to show that such a model is accurate and sufficient for radial distortion calibration, there has not been much analysis on the geometric/physical understanding of radial distortion from a camera calibration perspective. In this paper, we show using a thick-lens imaging model, that the variation of entrance pupil location as a function of incident image ray angle is directly responsible for radial distortion in captured images. Thus, unlike as proposed in the current state-of-the-art in camera calibration, radial distortion and entrance pupil movement are equivalent and need not be modeled together. By modeling only entrance pupil motion instead of radial distortion, we achieve two main benefits; first, we obtain comparable if not better pixel re-projection error than traditional methods; second, and more importantly, we directly back-project a radially distorted image pixel along the true image ray which formed it. Using a thick-lens setting, we show that such a back-projection is more accurate than the two-step method of undistorting an image pixel and then back-projecting it. We have applied this calibration method to the problem of generative depth-from-focus using focal stack to get accurate depth estimates.

****

A Linear Generalized Camera Calibration From Three Intersecting Reference Planes

Mai Nishimura, Shohei Nobuhara, Takashi Matsuyama, Shinya Shimizu, Kensaku Fujii; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2354-2362

This paper presents a new generalized (or ray-pixel, raxel) camera calibration algorithm for camera systems involving distortions by unknown refraction and reflection processes. The key idea is use of intersections of calibration planes, while conventional methods utilized collinearity constraints of points on the planes. We show that intersections of calibration planes can realize a simple linear algorithm, and that our method can be applied to any ray-distributions while conventional methods require knowing the ray-distribution class in advance. Evaluations using synthesized and real datasets demonstrate the performance of our method quantitatively and qualitatively.

****

Towards Pointless Structure From Motion: 3D Reconstruction and Camera Parameters From General 3D Curves

Irina Nurutdinova, Andrew Fitzgibbon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2363-2371

Modern structure from motion (SfM) remains dependent on point features to recover camera positions, meaning that reconstruction is severely hampered in low-texture environments, for example scanning a plain coffee cup on an uncluttered table. We show how 3D curves can be used to refine camera position estimation in challenging low-texture scenes. In contrast to previous work, we allow the curves to be partially observed in all images, meaning that for the first time, curve-based SfM can be demonstrated in realistic scenes. The algorithm is based on bundle adjustment, so needs an initial estimate, but even a poor estimate from a few point correspondences can be substantially improved by including curves, suggesting that this method would benefit many existing systems.

****

Attributed Grammars for Joint Estimation of Human Attributes, Part and Pose

Seyoung Park, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2372-2380

In this paper, we are interested in developing compositional models to explicit representing pose, parts and attributes and tackling the tasks of attribute recognition, pose estimation and part localization jointly. This is different from the recent trend of using CNN-based approaches for training and testing on these tasks separately with a large amount of data. Conventional attribute models typically use a large number of region-based attribute classifiers on parts of pre-trained pose estimator without explicitly detecting the object or its parts, or considering the correlations between attributes. In contrast, our approach jointly represents both the object parts and their semantic attributes within a unified compositional hierarchy. We apply our attributed grammar model to the task of human parsing by simultaneously performing part localization and attribute recognition. We show our modeling helps performance improvements on pose-estimation task and also outperforms on other existing methods on attribute prediction task.

****

Real-Time Pose Estimation Piggybacked on Object Detection

Roman Juranek, Adam Herout, Marketa Dubska, Pavel Zemcik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2381-2389

We present an object detector coupled with pose estimation directly in a single compact and simple model, where the detector shares extracted image features with the pose estimator. The output of the classification of each candidate window consists of both object score and likelihood map of poses. This extension introduces negligible overhead during detection so that the detector is still capable of real time operation. We evaluated the proposed approach on the problem of vehicle detection. We used existing datasets with viewpoint/pose annotation (WCVP, 3D objects, KITTI). Besides that, we collected a new traffic surveillance dataset COD20k which fills certain gaps of the existing datasets and we make it public. The experimental results show that the proposed approach is comparable with state-of-the-art approaches in terms of accuracy, but it is considerably faster - easily operating in real time (Matlab with C++ code). The source codes and the collected COD20k dataset are made public along with the paper.

****

Understanding and Predicting Image Memorability at a Large Scale

Aditya Khosla, Akhil S. Raju, Antonio Torralba, Aude Oliva; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2390-2398

Progress in estimating visual memorability has been limited by the small scale and lack of variety of benchmark data. Here, we introduce a novel experimental procedure to objectively measure human memory, building the largest annotated image memorability dataset to date (with 60,000 labeled images from a diverse array of sources). Using Convolutional Neural Networks (CNNs), we show that fine-tuned deep features outperform all other features by a large margin, reaching a rank correlation of 0.64, near human consistency (0.68). Analysis of the responses of the high-level CNN layers shows which objects and regions are positively, and negatively, correlated with memorability, allowing us to create memorability maps for each image and provide a concrete method to perform image memorability manipulation. This work demonstrates that one can now robustly estimate the memorability of images from many different classes, positioning memorability and deep memorability features as prime candidates to estimate the utility of information for cognitive systems.

****

Multiple Granularity Descriptors for Fine-Grained Categorization

Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, Zheng Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2399-2406

Fine-grained categorization, which aims to distinguish subordinate-level categories such as bird species or dog breeds, is an extremely challenging task. This is due to two main issues: how to localize discriminative regions for recognition and how to learn sophisticated features for representation. Neither of them is easy to handle if there is insufficient labeled data. We leverage the fact that a subordinate-level object already has other labels in its ontology tree. These "free" labels can be used to train a series of CNN-based classifiers, each specialized at one grain level. The internal representations of these networks have different region of interests, allowing the construction of multi-grained descriptors that encode informative and discriminative features covering all the grain levels. Our multiple granularity framework can be learned with the weakest supervision, requiring only image-level label and avoiding the use of labor-intensive bounding box or part annotations. Experimental results on three challenging fine-grained image datasets demonstrate that our approach outperforms state-of-the-art algorithms, including those requiring strong labels.

****

Guiding the Long-Short Term Memory Model for Image Caption Generation

Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2407-2415

In this work we focus on the problem of image caption generation. We propose an extension of the long short term memory (LSTM) model, which we coin gLSTM for short. In particular, we add semantic information extracted from the image as ext

ra input to each unit of the LSTM block, with the aim of guiding the model towards solutions that are more tightly coupled to the image content. Additionally, we explore different length normalization strategies for beam search to avoid bias towards short sentences. On various benchmark datasets such as Flickr8K, Flickr30K and MS COCO, we obtain results that are on par with or better than the current state-of-the-art.
****
Just Noticeable Differences in Visual Attributes
Aron Yu, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2416-2424
We explore the problem of predicting "just noticeable differences" in a visual attribute. While some pairs of images have a clear ordering for an attribute (e.g., A is more sporty than B), for others the difference may be indistinguishable to human observers. However, existing relative attribute models are unequipped to infer partial orders on novel data.   Attempting to map relative attribute ranks to equality predictions is non-trivial, particularly since the span of indistinguishable pairs in attribute space may vary in different parts of the feature space. We develop a Bayesian local learning strategy to infer when images are indistinguishable for a given attribute. On the UT-Zap50K shoes and LFW-10 faces datasets, we outperform a variety of alternative methods. In addition, we show the practical impact on fine-grained visual search.
****
VQA: Visual Question Answering
Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425-2433
We propose the task of free-form and open-ended Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing 0.25M images, 0.76M questions, and 10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines for VQA are provided and compared with human performance.
****
Localize Me Anywhere, Anytime: A Multi-Task Point-Retrieval Approach
Guoyu Lu, Yan Yan, Li Ren, Jingkuan Song, Nicu Sebe, Chandra Kambhamettu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2434-2442
Image-based localization is an essential complement to GPS localization. Current image-based localization methods are based on either 2D-to-3D or 3D-to-2D to find the correspondences, which ignore the real scene geometric attributes. The main contribution of our paper is that we use a 3D model reconstructed by a short video as the query to realize 3D-to-3D localization under a multi-task point retrieval framework. Firstly, the use of a 3D model as the query enables us to efficiently select location candidates. Furthermore, the reconstruction of 3D model exploits the correlations among different images, based on the fact that images captured from different views for SfM share information through matching features. By exploring shared information (matching features) across multiple related tasks (images of the same scene captured from different views), the visual feature's view-invariance property can be improved in order to get to a higher point retrieval accuracy. More specifically, we use multi-task point retrieval framework to explore the relationship between descriptors and the 3D points, which extracts the discriminant points for more accurate 3D-to-3D correspondences retrieval. We further apply multi-task learning (MTL) retrieval approach on thermal image

s to prove that our MTL retrieval framework also provides superior performance for the thermal domain. This application is exceptionally helpful to cope with the localization problem in an environment with limited light sources.
****

Dense Optical Flow Prediction From a Static Image
Jacob Walker, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2443-2451
Given a scene, what is going to move, and in what direction will it move? Such a question could be considered a non-semantic form of action prediction. In this work, we present a convolutional neural network (CNN) based approach for motion prediction. Given a static image, this CNN predicts the future motion of each and every pixel in the image in terms of optical flow. Our CNN model leverages the data in tens of thousands of realistic videos to train our model. Our method relies on absolutely no human labeling and is able to predict motion based on the context of the scene. Because our CNN model makes no assumptions about the underlying scene, it can predict future optical flow on a diverse set of scenarios. We outperform all previous approaches by large margins.
****

Unsupervised Domain Adaptation for Zero-Shot Learning
Elyor Kodirov, Tao Xiang, Zhenyong Fu, Shaogang Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2452-2460
Zero-shot learning (ZSL) can be considered as a special case of transfer learning where the source and target domains have different tasks/label spaces and the target domain is unlabelled, providing little guidance for the knowledge transfer. A ZSL method typically assumes that the two domains share a common semantic representation space, where a visual feature vector extracted from an image/video can be projected/embedded using a projection function. Existing approaches learn the projection function from the source domain and apply it without adaptation to the target domain. They are thus based on naive knowledge transfer and the learned projections are prone to the domain shift problem. In this paper a novel ZSL method is proposed based on unsupervised domain adaptation. Specifically, we formulate a novel regularised sparse coding framework which uses the target domain class labels' projections in the semantic space to regularise the learned target domain projection thus effectively overcoming the projection domain shift problem. Extensive experiments on four object and action recognition benchmark datasets show that the proposed ZSL method significantly outperforms the state-of-the-arts.
****

Visual Madlibs: Fill in the Blank Description Generation and Question Answering
Licheng Yu, Eunbyung Park, Alexander C. Berg, Tamara L. Berg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2461-2469
In this paper, we introduce a new dataset consisting of 360,001 focused natural language descriptions for 10,738 images. This dataset, the Visual Madlibs dataset, is collected using automatically produced fill-in-the-blank templates designed to gather targeted descriptions about: people and objects, their appearances, activities, and interactions, as well as inferences about the general scene or its broader context. We provide several analyses of the Visual Madlibs dataset and demonstrate its applicability to two new description generation tasks: focused description generation, and multiple-choice question-answering for images. Experiments using joint-embedding and deep learning methods show promising results on these tasks.
****

Actions and Attributes From Wholes and Parts
Georgia Gkioxari, Ross Girshick, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2470-2478
We investigate the importance of parts for the tasks of action and attribute classification. We develop a part-based approach by leveraging convolutional network features inspired by recent advances in computer vision. Our part detectors are a deep version of poselets and capture parts of the human body under a distinct set of poses. For the tasks of action and attribute classification, we train h

olistic convolutional neural networks and show that adding parts leads to top-performing results for both tasks. We observe that for deeper networks parts are less significant. In addition, we demonstrate the effectiveness of our approach when we replace an oracle person detector, as is the default in the current evaluation protocol for both tasks, with a state-of-the-art person detection system.

****

DeepBox: Learning Objectness With Convolutional Networks

Weicheng Kuo, Bharath Hariharan, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2479-2487

Existing object proposal approaches use primarily bottom-up cues to rank proposals, while we believe that "objectness" is in fact a high level construct. We argue for a data-driven, semantic approach for ranking object proposals. Our framework, which we call DeepBox, uses convolutional neural networks (CNNs) to rerank proposals from a bottom-up method. We use a novel four-layer CNN architecture that is as good as much larger networks on the task of evaluating objectness while being much faster. We show that DeepBox significantly improves over the bottom-up ranking, achieving the same recall with 500 proposals as achieved by bottom-up methods with 2000. This improvement generalizes to categories the CNN has never seen before and leads to a 4.5-point gain in detection mAP. Our implementation achieves this performance while running at 260 ms per image.

****

Active Object Localization With Deep Reinforcement Learning

Juan C. Caicedo, Svetlana Lazebnik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2488-2496

We present an active detection model for localizing objects in scenes. The model is class-specific and allows an agent to focus attention on candidate regions for identifying the correct location of a target object. This agent learns to deform a bounding box using simple transformation actions, with the goal of determining the most specific location of target objects following top-down reasoning. The proposed localization agent is trained using deep reinforcement learning, and evaluated on the Pascal VOC 2007 dataset. We show that agents guided by the proposed model are able to localize a single instance of an object after analyzing only between 11 and 25 regions in an image, and obtain the best detection results among systems that do not use object proposals for object localization.

****

Scene-Domain Active Part Models for Object Representation

Zhou Ren, Chaohui Wang, Alan L. Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2497-2505

In this paper, we are interested in enhancing the expressivity and robustness of part-based models for object representation, in the common scenario where the training data are based on 2D images. To this end, we propose scene-domain active part models (SDAPM), which reconstruct and characterize the 3D geometric statistics between object's parts in 3D scene-domain by using 2D training data in the image-domain alone. And on top of this, we explicitly model and handle occlusions in SDAPM. Together with the developed learning and inference algorithms, such a model provides rich object descriptions, including 2D object and parts localization, 3D landmark shape and camera viewpoint, which offers an effective representation to various image understanding tasks, such as object and parts detection, 3D landmark shape and viewpoint estimation from images. Experiments on the above tasks show that SDAPM outperforms previous part-based models, and thus demonstrates the potential of the proposed technique.

****

A Unified Multiplicative Framework for Attribute Learning

Kongming Liang, Hong Chang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2506-2514

Attributes are mid-level semantic properties of objects. Recent research has shown that visual attributes can benefit many traditional learning problems in computer vision community. However, attribute learning is still a challenging problem as the attributes may not always be predictable directly from input images and the variation of visual attributes is sometimes large across categories. In thi s paper, we propose a unified multiplicative framework for attribute learning, which tackles the key problems. Specifically, images and category information are jointly projected into a shared feature space, where the latent factors are disentangled and multiplied for attribute prediction. The resulting attribute classifier is category-specific instead of being shared by all categories. Moreover, our method can leverage auxiliary data to enhance the predictive ability of attribute classifiers, reducing the effort of instance-level attribute annotation to some extent. Experimental results show that our method achieves superior performance on both instance-level and category-level attribute prediction. For zero-shot learning based on attributes, our method significantly improves the state-of-the-art performance on AwA dataset and achieves comparable performance on CUB dataset.

****

Contractive Rectifier Networks for Nonlinear Maximum Margin Classification

Senjian An, Munawar Hayat, Salman H. Khan, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2515-2523

To find the optimal nonlinear separating boundary with maximum margin in the input data space, this paper proposes Contractive Rectifier Networks (CRNs), wherein the hidden-layer transformations are restricted to be contraction mappings. The contractive constraints ensure that the achieved separating margin in the input space is larger than or equal to the separating margin in the output layer. The training of the proposed CRNs is formulated as a linear support vector machine (SVM) in the output layer, combined with two or more contractive hidden layers. Effective algorithms have been proposed to address the optimization challenges arising from contraction constraints. Experimental results on MNIST, CIFAR-10, CIFAR-100 and MIT-67 datasets demonstrate that the proposed contractive rectifier networks consistently outperform their conventional unconstrained rectifier network counterparts.

****

Augmenting Strong Supervision Using Web Data for Fine-Grained Categorization

Zhe Xu, Shaoli Huang, Ya Zhang, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2524-2532

We propose a new method for fine-grained object recognition that employs part-level annotations and deep convolutional neural networks (CNNs) in a unified framework. Although both schemes have been widely used to boost recognition performance, due to the difficulty in acquiring detailed part annotations, strongly supervised fine-grained datasets are usually too small to keep pace with the rapid evolution of CNN architectures. In this paper, we solve this problem by exploiting inexhaustible web data. The proposed method improves classification accuracy in two ways: more discriminative CNN feature representations are generated using a training set augmented by collecting a large number of part patches from weakly supervised web images; and more robust object classifiers are learned using a multi-instance learning algorithm jointly on the strong and weak datasets. Despite its simplicity, the proposed method delivers a remarkable performance improvement on the CUB200-2011 dataset compared to baseline part-based R-CNN methods, and achieves the highest accuracy on this dataset even in the absence of test image annotations.

****

Learning Like a Child: Fast Novel Visual Concept Learning From Sentence Descriptions of Images

Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, Alan L. Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2533-2541

In this paper, we address the task of learning novel visual concepts, and their interactions with other concepts, from a few images with sentence descriptions. Using linguistic context and visual features, our method is able to efficiently hypothesize the semantic meaning of new words and add them to its word dictionary so that they can be used to describe images which contain these novel concepts. Our method has an image captioning module based on the m-RNN model with severa l improvements. In particular, we propose a transposed weight sharing scheme, which not only improves performance on image captioning, but also makes the model more suitable for the novel concept learning task. We propose methods to prevent overfitting the new concepts. In addition, three novel concept datasets are constructed for this new task, and are publicly available on the project page. In the experiments, we show that our method effectively learns novel visual concepts from a few examples without disturbing the previously learned concepts. The project page is: http://www.stat.ucla.edu/ junhua.mao/projects/child_learning.html

****

Learning Common Sense Through Visual Abstraction

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, Devi Parikh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2542-2550

Common sense is essential for building intelligent machines. While some commonsense knowledge is explicitly stated in human-generated text and can be learnt by mining the web, much of it is unwritten. It is often unnecessary and even unnatural to write about commonsense facts. While unwritten, this commonsense knowledge is not unseen! The visual world around us is full of structure modeled by commonsense knowledge. Can machines learn common sense simply by observing our visual world? Unfortunately, this requires automatic and accurate detection of objects, their attributes, poses, and interactions between objects, which remain challenging problems. Our key insight is that while visual common sense is depicted in visual content, it is the semantic features that are relevant and not low-level pixel information. In other words, photorealism is not necessary to learn common sense. We explore the use of human-generated abstract scenes made from clipart for learning common sense. In particular, we reason about the plausibility of an interaction or relation between a pair of nouns by measuring the similarity of the relation and nouns with other relations and nouns we have seen in abstract scenes. We show that the commonsense knowledge we learn is complementary to what can be learnt from sources of text.

****

Domain Generalization for Object Recognition With Multi-Task Autoencoders

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2551-2559

The problem of domain generalization is to take knowledge acquired from a number of related domains, where training data is available, and to then successfully apply it to previously unseen domains. We propose a new feature learning algorithm, Multi-Task Autoencoder (MTAE), that provides good generalization performance for cross-domain object recognition. The algorithm extends the standard denoising autoencoder framework by substituting artificially induced corruption with naturally occurring inter-domain variability in the appearance of objects. Instead of reconstructing images from noisy versions, MTAE learns to transform the original image into analogs in multiple related domains. It thereby learns features that are robust to variations across domains. The learnt features are then used as inputs to a classifier. We evaluated the performance of the algorithm on benchmark image recognition datasets, where the task is to learn features from multiple datasets and to then predict the image label from unseen datasets. We found that (denoising) MTAE outperforms alternative autoencoder-based models as well as the current state-of-the-art algorithms for domain generalization.

****

Square Localization for Efficient and Accurate Object Detection

Cewu Lu, Yongyi Lu, Hao Chen, Chi-Keung Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2560-2568

The key contribution of this paper is the compact square object localization, which relaxes the exhaustive sliding window from testing all windows of different combinations of aspect ratios. Square object localization is category scalable. By using a binary search strategy, the number of scales to test is further reduced empirically to only O(log(minfH;Wg)) rounds of sliding CNNs, where H and W are respectively the image height and width. In the training phase, square CNN mod els and object co-presence priors are learned. In the testing phase, sliding CNN models are applied which produces a set of response maps that can be effectively filtered by the learned co-presence prior to output the final bounding boxes for localizing an object. We performed extensive experimental evaluation on the VOC 2007 and 2012 datasets to demonstrate that while efficient,square localization can output precise bounding boxes to improve the final detection result.

****

Box Aggregation for Proposal Decimation: Last Mile of Object Detection

Shu Liu, Cewu Lu, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2569-2577

Regions-with-convolutional-neural-network (RCNN) is now a commonly employed object detection pipeline. Its main steps, i.e., proposal generation and convolutional neural network (CNN) feature extraction, have been intensively investigated. We focus on the last step of the system to aggregate thousands of scored box proposals into final object prediction, which we call proposal decimation. We show this step can be enhanced with a very simple box aggregation function by considering statistical properties of proposals with respect to ground truth objects. Our method is with extremely light-weight computation, while it yields an improvement of 3.7% in mAP on PASCAL VOC 2007 test. We explain why it works using some statistics in this paper.

****

DeepProposal: Hunting Objects by Cascading Deep Convolutional Layers

Amir Ghodrati, Ali Diba, Marco Pedersoli, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2578-2586

In this paper we evaluate the quality of the activation layers of a convolutional neural network (CNN) for the generation of object proposals. We generate hypotheses in a sliding-window fashion over different activation layers and show that the final convolutional layers can find the object of interest with high recall but poor localization due to the coarseness of the feature maps. Instead, the first layers of the network can better localize the object of interest but with a reduced recall. Based on this observation we design a method for proposing object locations that is based on CNN features and that combines the best of both worlds. We build an inverse cascade that, going from the final to the initial convolutional layers of the CNN, selects the most promising object locations and refines their boxes in a coarse-to-fine manner. The method is efficient, because i) it uses the same features extracted for detection, ii) it aggregates features using integral images, and iii) it avoids a dense evaluation of the proposals due to the inverse coarse-to-fine cascade. The method is also accurate; it outperforms most of the previously proposed object proposals approaches and when plugged into a CNN-based detector produces state-of-the-art detection performance.

****

Semantic Segmentation With Object Clique Potential

Xiaojuan Qi, Jianping Shi, Shu Liu, Renjie Liao, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2587-2595

In this paper, we propose an object clique potential for semantic segmentation. Our object clique potential addresses the misclassified object-part issues arising in solutions based on fully-connected networks. Our object clique set, compared to that yielded from segment-proposal-based approaches, is with a significantly small size, making our method consume notably less computation. Regarding system design and model formation, our object clique potential can be regarded as a functionally complement to local-appearance-based CRF models and works in synergy with these effective approaches for further performance improvement. Extensive experiments verify our method.

****

Automatic Concept Discovery From Parallel Text and Visual Corpora

Chen Sun, Chuang Gan, Ram Nevatia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2596-2604

Humans connect language and vision to perceive the world. How to build a similar connection for computers? One possible way is via visual concepts, which are te xt terms that relate to visually discriminative entities. We propose an automatic visual concept discovery algorithm using parallel text and visual corpora; it filters text terms based on the visual discriminative power of the associated images, and groups them into concepts using visual and semantic similarities. We illustrate the applications of the discovered concepts using bidirectional image and sentence retrieval task and image tagging task, and show that the discovered concepts not only outperform several large sets of manually selected concepts significantly, but also achieves the state-of-the-art performance in the retrieval task.

****

Simpler Non-Parametric Methods Provide as Good or Better Results to Multiple-Instance Learning

Ragav Venkatesan, Parag Chandakkar, Baoxin Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2605-2613

Multiple-instance learning (MIL) is a unique learning problem in which training data labels are available only for collections of objects (called bags) instead of individual objects (called instances). A plethora of approaches have been developed to solve this problem in the past years. Popular methods include the diverse density, MILIS and DD-SVM. While having been widely used, these methods, particularly those in computer vision have attempted fairly sophisticated solutions to solve certain unique and particular configurations of the MIL space. In this paper, we analyze the MIL feature space using modified versions of traditional non-parametric techniques like the Parzen window and k-nearest-neighbour, and develop a learning approach employing distances to k-nearest neighbours of a point in the feature space. We show that these methods work as well, if not better than most recently published methods on benchmark datasets. We compare and contrast our analysis with the well-established diverse-density approach and its variants in recent literature, using benchmark datasets including the Musk, Andrews' and Corel datasets, along with a diabetic retinopathy pathology diagnosis dataset. Experimental results demonstrate that, while enjoying an intuitive interpretation and supporting fast learning, these method have the potential of delivering improved performance even for complex data arising from real-world applications.

****

Monocular Object Instance Segmentation and Depth Ordering With CNNs

Ziyu Zhang, Alexander G. Schwing, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2614-2622

In this paper we tackle the problem of instance-level segmentation and depth ordering from a single monocular image. Towards this goal, we take advantage of convolutional neural nets and train them to directly predict instance-level segmentations where the instance ID encodes the depth ordering within image patches. To provide a coherent single explanation of an image we develop a Markov random field which takes as input the predictions of convolutional neural nets applied at overlapping patches of different resolutions, as well as the output of a connected component algorithm. It aims to predict accurate instance-level segmentation and depth ordering. We demonstrate the effectiveness of our approach on the challenging KITTI benchmark and show good performance on both tasks.

****

Multimodal Convolutional Neural Networks for Matching Image and Sentence

Lin Ma, Zhengdong Lu, Lifeng Shang, Hang Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2623-2631

In this paper, we propose multimodal convolutional neural networks (m-CNNs) for matching image and sentence. Our m-CNN provides an end-to-end framework with convolutional architectures to exploit image representation, word composition, and the matching relations between the two modalities. More specifically, it consists of one image CNN encoding the image content and one matching CNN modeling the joint representation of image and sentence. The matching CNN composes different semantic fragments from words and learns the inter-modal relations between image and the composed fragments at different levels, thus fully exploit the matching relations between image and sentence. Experimental results demonstrate that the proposed m-CNNs can effectively capture the information necessary for image and sentence matching. More specifically, our proposed m-CNNs significantly outperform the state-of-the-art approaches for bidirectional image and sentence retrieval on the Flickr8K and Flickr30K datasets.

****

Structural Kernel Learning for Large Scale Multiclass Object Co-Detection

Zeeshan Hayder, Xuming He, Mathieu Salzmann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2632-2640

Exploiting contextual relationships across images has recently proven key to improve object detection. The resulting object co-detection algorithms, however, fail to exploit the correlations between multiple classes and, for scalability reasons are limited to modeling object instance similarity with relatively low-dimensional hand-crafted features. Here, we address the problem of multiclass object co-detection for large scale datasets. To this end, we formulate co-detection as the joint multiclass labeling of object candidates obtained in a class-independent manner. To exploit the correlations between objects, we build a fully-connected CRF on the candidates, which explicitly incorporates both geometric layout relations across object classes and similarity relations across multiple images. We then introduce a structural boosting algorithm that lets us exploits rich, high-dimensional deep network features to learn object similarity within our fully-connected CRF. Our experiments on PASCAL VOC 2007 and 2012 evidences the benefits of our approach over object detection with RCNN, single-image CRF methods and state-of-the-art co-detection algorithms.

****

Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, Svetlana Lazebnik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2641-2649

The Flickr30k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains linking mentions of the same entities in images, as well as 276k manually annotated bounding boxes corresponding to each entity. Such annotation is essential for continued progress in automatic image description and grounded language understanding. We present experiments demonstrating the usefulness of our annotations for text-to-image reference resolution, or the task of localizing textual entity mentions in an image, and for bidirectional image-sentence retrieval. These experiments confirm that we can further improve the accuracy of state-of-the-art retrieval methods by training with explicit region-to-phrase correspondence, but at the same time, they show that accurately inferring this correspondence given an image and a caption remains really challenging.

****

Predicting Depth, Surface Normals and Semantic Labels With a Common Multi-Scale Convolutional Architecture

David Eigen, Rob Fergus; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2650-2658

In this paper we address three different computer vision tasks using a single basic architecture: depth prediction, surface normal estimation, and semantic labeling. We use a multiscale convolutional network that is able to adapt easily to each task using only small modifications, regressing from the input image to the output map directly. Our method progressively refines predictions using a sequence of scales, and captures many image details without any superpixels or low-level segmentation. We achieve state-of-the-art performance on benchmarks for all three tasks.

****

AttentionNet: Aggregating Weak Directions for Accurate Object Detection

Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S. Paek, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2659-2667

We present a novel detection method using a deep convolutional neural network (CNN), named AttentionNet. We cast an object detection problem as an iterative classification problem, which is the most suitable form of a CNN. AttentionNet provides quantized weak directions pointing a target object and the ensemble of iterative predictions from AttentionNet converges to an accurate object boundary box. Since AttentionNet is a unified network for object detection, it detects objects without any separated models from the object proposal to the post bounding-box regression. We evaluate AttentionNet by a human detection task and achieve the state-of-the-art performance of 65% (AP) on PASCAL VOC 2007/2012 with an 8-layered architecture only.

****

Common Subspace for Model and Similarity: Phrase Learning for Caption Generation From Images

Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, Tatsuya Harada; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2668-2676

Generating captions to describe images is a fundamental problem that combines computer vision and natural language processing. Recent works focus on descriptive phrases, such as "a white dog" to explain the visual composites of an input image. The phrases can not only express objects, attributes, events, and their relations but can also reduce visual complexity. A caption for an input image can be generated by connecting estimated phrases using a grammar model. However, because phrases are combinations of various words, the number of phrases is much larger than the number of single words. Consequently, the accuracy of phrase estimation suffers from too few training samples per phrase. In this paper, we propose a novel phrase-learning method: Common Subspace for Model and Similarity (CoSMoS). In order to overcome the shortage of training samples, CoSMoS obtains a subspace in which (a) all feature vectors associated with the same phrase are mapped as mutually close, (b) classifiers for each phrase are learned, and (c) training samples are shared among co-occurring phrases. Experimental results demonstrate that our system is more accurate than those in earlier work and that the accuracy increases when the dataset from the web increases.

****

3D-Assisted Feature Synthesis for Novel Views of an Object

Hao Su, Fan Wang, Eric Yi, Leonidas J. Guibas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2677-2685

Comparing two images from different views has been a long-standing challenging problem in computer vision, as visual features are not stable under large view point changes. In this paper, given a single input image of an object, we synthesize its features for other views, leveraging an existing modestly-sized 3D model collection of related but not identical objects.To accomplish this, we study the relationship of image patches between different views of the same object, seeking what we call surrogate patches -- patches in one view whose feature content predicts well the features of a patch in another view. Based upon these surrogate relationships, we can create feature sets for all views of the latent object on a per patch basis, providing us an augmented multi-view representation of the object. We provide theoretical and empirical analysis of the feature synthesis process, and evaluate the augmented features in fine-grained image retrieval/recognition and instance retrieval tasks. Experimental results show that our synthesized features do enable view-independent comparison between images and perform significantly better than other traditional approaches in this respect.

****

Render for CNN: Viewpoint Estimation in Images Using CNNs Trained With Rendered 3D Model Views

Hao Su, Charles R. Qi, Yangyan Li, Leonidas J. Guibas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2686-2694

Object viewpoint estimation from 2D images is an essential task in computer vision. However, two issues hinder its progress: scarcity of training data with viewpoint annotations, and a lack of powerful features. Inspired by the growing availability of 3D models, we propose a framework to address both issues by combinin g render-based image synthesis and CNNs (Convolutional Neural Networks). We believe that 3D models have the potential in generating a large number of images of high variation, which can be well exploited by deep CNN with a high learning capacity. Towards this goal, we propose a scalable and overfit-resistant image synthesis pipeline, together with a novel CNN specifically tailored for the viewpoint estimation task. Experimentally, we show that the viewpoint estimation from our pipeline can significantly outperform state-of-the-art methods on PASCAL 3D+ benchmark.

****

Lost Shopping! Monocular Localization in Large Indoor Spaces

Shenlong Wang, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2695-2703

In this paper we propose a novel approach to localization in very large indoor spaces (i.e., 200+ store shopping malls) that takes   a single image and a floor plan of the environment as input. We formulate the localization problem as inference in a Markov random field, which jointly reasons about text detection (localizing shop's names in the image with precise bounding boxes), shop facade segmentation, as well as   camera's rotation and translation within the entire shopping mall. The power of our approach is that it does not use any prior information about appearance and instead exploits text detections corresponding to the shop names. This makes our method applicable to a variety of domains and robust to store appearance variation across countries, seasons, and illumination conditions. We demonstrate the performance of our approach in a new dataset we collected of two very large shopping malls, and show the power of holistic reasoning.

****

Camera Pose Voting for Large-Scale Image-Based Localization

Bernhard Zeisl, Torsten Sattler, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2704-2712

Image-based localization approaches aim to determine the camera pose from which an image was taken. Finding correct 2D-3D correspondences between query image features and 3D points in the scene model becomes harder as the size of the model increases. Current state-of-the-art methods therefore combine elaborate matching schemes with camera pose estimation techniques that are able to handle large fractions of wrong matches. In this work we study the benefits and limitations of spatial verification compared to appearance-based filtering. We propose a voting-based pose estimation strategy that exhibits O(n) complexity in the number of matches and thus facilitates to consider much more matches than previous approaches - whose complexity grows at least quadratically. This new outlier rejection formulation enables us to evaluate pose estimation for 1-to-many matches and to surpass the state-of-the-art. At the same time, we show that using more matches does not automatically lead to a better performance.

****

MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking

Thibaut Durand, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2713-2721

In this work, we propose a novel Weakly Supervised Learning (WSL) framework dedicated to learn discriminative part detectors from images annotated with a global label. Our WSL method encompasses three main contributions. Firstly, we introduce a new structured output latent variable model, Minimum mAximum lateNt sTRuctural SVM (MANTRA), which prediction relies on a pair of latent variables: h^+ (resp. h^-) provides positive (resp. negative) evidence for a given output y. Secondly, we instantiate MANTRA for two different visual recognition tasks: multi-class classification and ranking. For ranking, we propose efficient solutions to exactly solve the inference and the loss-augmented problems. Finally, extensive experiments highlight the relevance of the proposed method: MANTRA outperforms state-of-the art results on five different datasets.

****

DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving

Chenyi Chen, Ari Seff, Alain Kornhauser, Jianxiong Xiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2722-2730

Today, there are two major paradigms for vision-based autonomous driving systems: mediated perception approaches that parse an entire scene to make a driving decision, and behavior reflex approaches that directly map an input image to a driving action by a regressor. In this paper, we propose a third paradigm: a direct perception approach to estimate the affordance for driving. We propose to map an input image to a small number of key perception indicators that directly relate to the affordance of a road/traffic state for driving. Our representation provides a set of compact yet complete descriptions of the scene to enable a simple controller to drive autonomously. Falling in between the two extremes of mediated perception and behavior reflex, we argue that our direct perception representation provides the right level of abstraction. To demonstrate this, we train a deep Convolutional Neural Network using recording from 12 hours of human driving in a video game and show that our model can work well to drive a car in a very diverse set of virtual environments. We also train a model for car distance estimation on the KITTI dataset. Results show that our direct perception approach can generalize well to real driving images. Source code and data are available on our project website.

****

Active Transfer Learning With Zero-Shot Priors: Reusing Past Datasets for Future Tasks

Efstratios Gavves, Thomas Mensink, Tatiana Tommasi, Cees G. M. Snoek, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2731-2739

How can we reuse existing knowledge, in the form of available datasets, when solving a new and apparently unrelated target task from a set of unlabeled data? In this work we make a first contribution to answer this question in the context of image classification. We frame this quest as an active learning problem and use zero-shot classifiers to guide the learning process by linking the new task to the the existing classifiers. By revisiting the dual formulation of adaptive SVM, we reveal two basic conditions to choose greedily only the most relevant samples to be annotated. On this basis we propose an effective active learning algorithm which learns the best possible target classification model with minimum human labeling effort. Extensive experiments on two challenging datasets show the value of our approach compared to the state-of-the-art active learning methodologies, as well as its potential to reuse past datasets with minimal effort for future tasks.

****

HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, Yizhou Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2740-2748

In image classification, visual separability between different object categories is highly uneven, and some categories are more difficult to distinguish than others. Such difficult categories demand more dedicated classifiers. However, existing deep convolutional neural networks (CNN) are trained as flat N-way classifiers, and few efforts have been made to leverage the hierarchical structure of categories. In this paper, we introduce hierarchical deep CNNs (HD-CNNs) by embedding deep CNNs into a category hierarchy. An HD-CNN separates easy classes using a coarse category classifier while distinguishing difficult classes using fine category classifiers. During HD-CNN training, component-wise pretraining is followed by global finetuning with a multinomial logistic loss regularized by a coarse category consistency term. In addition, conditional executions of fine category classifiers and layer parameter compression make HD-CNNs scalable for large-scale visual recognition. We achieve state-of-the-art results on both CIFAR100 and large-scale ImageNet 1000-class benchmark datasets. In our experiments, we build up three different HD-CNNs and they lower the top-1 error of the standard CNNs by 2.65%, 3.1% and 1.1%, respectively

****

Learning The Structure of Deep Convolutional Networks

Jiashi Feng, Trevor Darrell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2749-2757

In this work, we develop a novel method for automatically learning aspects of the structure of a deep model, in order to improve its performance, especially when labeled training data are scarce. We propose a new convolutional neural network model with the Indian Buffet Process (IBP) prior, termed ibpCNN. The ibpCNN automatically adapts its structure to provided training data, achieves an optimal balance among model complexity, data fidelity and training loss, and thus offers better generalization performance. The proposed ibpCNN captures complicated data distribution in an unsupervised generative way. Therefore, ibpCNN can exploit unlabeled data -- which can be collected at low cost -- to learn its structure. After determining the structure, ibpCNN further learns its parameters according to specified tasks, in an end-to-end fashion, and produces discriminative yet compact representations. We evaluate the performance of ibpCNN, on fully- and semi-supervised image classification tasks; ibpCNN surpasses standard CNN models on benchmark datasets, with much smaller size and higher efficiency.

****

FlowNet: Learning Optical Flow With Convolutional Networks

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2758-2766

Convolutional neural networks (CNNs) have recently been very successful in a variety of computer vision tasks, especially on those linked to recognition. Optical flow estimation has not been among the tasks CNNs succeeded at. In this paper we construct CNNs which are capable of solving the optical flow estimation problem as a supervised learning task. We propose and compare two architectures: a generic architecture and another one including a layer that correlates feature vectors at different image locations. Since existing ground truth data sets are not sufficiently large to train a CNN, we generate a large synthetic Flying Chairs dataset. We show that networks trained on this unrealistic data still generalize very well to existing datasets such as Sintel and KITTI, achieving competitive accuracy at frame rates of 5 to 10 fps.

****

Learning Semi-Supervised Representation Towards a Unified Optimization Framework for Semi-Supervised Learning

Chun-Guang Li, Zhouchen Lin, Honggang Zhang, Jun Guo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2767-2775

State of the art approaches for Semi-Supervised Learning (SSL) usually follow a two-stage framework -- constructing an affinity matrix from the data and then propagating the partial labels on this affinity matrix to infer those unknown labels. While such a two-stage framework has been successful in many applications, solving two subproblems separately only once is still suboptimal because it does not fully exploit the correlation between the affinity and the labels. In this paper, we formulate the two stages of SSL into a unified optimization framework, which learns both the affinity matrix and the unknown labels simultaneously. In the unified framework, both the given labels and the estimated labels are used to learn the affinity matrix and to infer the unknown labels. We solve the unified optimization problem via an alternating direction method of multipliers combined with label propagation. Extensive experiments on a synthetic data set and several benchmark data sets demonstrate the effectiveness of our approach.

****

Context-Guided Diffusion for Label Propagation on Graphs

Kwang In Kim, James Tompkin, Hanspeter Pfister, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2776-2784

Existing approaches for diffusion on graphs, e.g., for label propagation, are mainly focused on isotropic diffusion, which is induced by the commonly-used graph

Laplacian regularizer. Inspired by the success of diffusivity tensors for anisotropic diffusion in image processing, we presents anisotropic diffusion on graphs and the corresponding label propagation algorithm. We develop positive definite diffusivity operators on the vector bundles of Riemannian manifolds, and discretize them to diffusivity operators on graphs. This enables us to easily define new robust diffusivity operators which significantly improve semi-supervised learning performance over existing diffusion algorithms.

****

Learning to Rank Based on Subsequences

Basura Fernando, Efstratios Gavves, Damien Muselet, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2785-2793

We present a supervised learning to rank algorithm that effectively orders images by exploiting the structure in image sequences. Most often in the supervised learning to rank literature, ranking is approached either by analysing pairs of images or by optimizing a list-wise surrogate loss function on full sequences. In this work we propose MidRank, which learns from moderately sized sub-sequences instead. These sub-sequences contain useful structural ranking information that leads to better learnability during training and better generalization during testing. By exploiting sub-sequences, the proposed MidRank improves ranking accuracy considerably on an extensive array of image ranking applications and datasets.

****

Unsupervised Learning of Visual Representations Using Videos

Xiaolong Wang, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2794-2802

Is strong supervision necessary for learning a good visual representation? Do we really need millions of semantically-labeled images to train a Convolutional Neural Network (CNN)? In this paper, we present a simple yet surprisingly powerful approach for unsupervised learning of CNN. Specifically, we use hundreds of thousands of unlabeled videos from the web to learn visual representations. Our key idea is that visual tracking provides the supervision. That is, two patches connected by a track should have similar visual representation in deep feature space since they probably belong to same object or object part. We design a Siamese-triplet network with a ranking loss function to train this CNN representation. Without using a single image from ImageNet, just using 100K unlabeled videos and the VOC 2012 dataset, we train an ensemble of unsupervised networks that achieves 52% mAP (no bounding box regression). This performance comes tantalizingly close to its ImageNet-supervised counterpart, an ensemble which achieves a mAP of 54.4%. We also show that our unsupervised network can perform competitively in other tasks such as surface-normal estimation.

****

A Nonparametric Bayesian Approach Toward Stacked Convolutional Independent Component Analysis

Sotirios P. Chatzis, Dimitrios Kosmopoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2803-2811

Unsupervised feature learning algorithms based on convolutional formulations of independent components analysis (ICA) have been demonstrated to yield state-of-the-art results in several action recognition benchmarks. However, existing approaches do not allow for the number of latent components (features) to be automatically inferred from the data in an unsupervised manner. This is a significant disadvantage of the state-of-the-art, as it results in considerable burden imposed on researchers and practitioners, who must resort to tedious cross-validation procedures to obtain the optimal number of latent features. To resolve these issues, in this paper we introduce a convolutional nonparametric Bayesian sparse ICA architecture for overcomplete feature learning from high-dimensional data. Our method utilizes an Indian buffet process prior to facilitate inference of the appropriate number of latent features under a hybrid variational inference algorithm, scalable to massive datasets. As we show, our model can be naturally used to obtain deep unsupervised hierarchical feature extractors, by greedily stacking successive model layers, similar to existing approaches. In addition, inference for this model is completely heuristics-free; thus, it obviates the need of tedious parameter tuning, which is a major challenge most deep learning approaches are faced with. We evaluate our method on several action recognition benchmarks, and exhibit its advantages over the state-of-the-art.

****

Robust Principal Component Analysis on Graphs

Nauman Shahid, Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, Pierre Vandergheynst; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2812-2820

Principal Component Analysis (PCA) is the most widely used tool for linear dimensionality reduction and clustering. Still it is highly sensitive to outliers and does not scale well with respect to the number of data samples. Robust PCA solves the first issue with a sparse penalty term. The second issue can be handled with the matrix factorization model, which is however non-convex. Besides, PCA based clustering can also be enhanced by using a graph of data similarity. In this article, we introduce a new model called 'Robust PCA on Graphs' which incorporates spectral graph regularization into the Robust PCA framework. Our proposed model benefits from 1) the robustness of principal components to occlusions and missing values, 2) enhanced low-rank recovery, 3) improved clustering property due to the graph smoothness assumption on the low-rank matrix, and 4) convexity of the resulting optimization problem. Extensive experiments on 8 benchmark, 3 video and 2 artificial datasets with corruptions clearly reveal that our model outperforms 10 other state-of-the-art models in its clustering and low-rank recovery tasks.

****

Projection Bank: From High-Dimensional Data to Medium-Length Binary Codes

Li Liu, Mengyang Yu, Ling Shao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2821-2829

Recently, very high-dimensional feature representations, e.g., Fisher Vector, have achieved excellent performance for visual recognition and retrieval. However, these lengthy representations always cause extremely heavy computational and storage costs and even become unfeasible in some large-scale applications. A few existing techniques can transfer very high-dimensional data into binary codes, but they still require the reduced code length to be relatively long to maintain acceptable accuracies. To target a better balance between computational efficiency and accuracies, in this paper, we propose a novel embedding method called Binary Projection Bank (BPB), which can effectively reduce the very high-dimensional representations to medium-dimensional binary codes without sacrificing accuracies. Instead of using conventional single linear or bilinear projections, the proposed method learns a bank of small projections via the max-margin constraint to optimally preserve the intrinsic data similarity. We have systematically evaluated the proposed method on three datasets: Flickr 1M, ILSVR2010 and UCF101, showing competitive retrieval and recognition accuracies compared with state-of-the-art approaches, but with a significantly smaller memory footprint and lower coding complexity.

****

Robust Optimization for Deep Regression

Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, Nassir Navab; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2830-2838

Convolutional Neural Networks (ConvNets) have successfully contributed to improve the accuracy of regression-based methods for computer vision tasks such as human pose estimation, landmark localization, and object detection. The network optimization has been usually performed with L2 loss and without considering the impact of outliers on the training process, where an outlier in this context is defined by a sample estimation that lies at an abnormal distance from the other training sample estimations in the objective space. In this work, we propose a regression model with ConvNets that achieves robustness to such outliers by minimizing Tukey's biweight function, an M-estimator robust to outliers, as the loss fu nction for the ConvNet. In addition to the robust loss, we introduce a coarse-to-fine model, which processes input images of progressively higher resolutions for improving the accuracy of the regressed values. In our experiments, we demonstrate faster convergence and better generalization of our robust loss function for the tasks of human pose estimation and age estimation from face images. We also show that the combination of the robust loss function with the coarse-to-fine model produces comparable or better results than current state-of-the-art approaches in four publicly available human pose estimation datasets.

****

Multi-Class Multi-Annotator Active Learning With Robust Gaussian Process for Visual Recognition

Chengjiang Long, Gang Hua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2839-2847

Active learning is an effective way to relieve the tedious work of manual annotation in many applications of visual recognition. However, less research attention has been focused on multi-class active learning. In this paper, we propose a novel Gaussian process classifier model with multiple annotators for multi-class visual recognition. Expectation propagation (EP) is adopted for efficient approximate Bayesian inference of our probabilistic model for classification. Based on the EP approximation inference, a generalized Expectation Maximization (GEM) algorithm is derived to estimate both the parameters for instances and the quality of each individual annotator. Also, we incorporate the idea of reinforcement learning to actively select both the informative samples and the high-quality annotators, which better explores the trade-off between exploitation and exploration. The experiments clearly demonstrate the efficacy of the proposed model.

****

Maximum-Margin Structured Learning With Deep Networks for 3D Human Pose Estimation

Sijin Li, Weichen Zhang, Antoni B. Chan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2848-2856

This paper focuses on structured-output learning using deep neural networks for 3D human pose estimation from monocular images. Our network takes an image and 3D pose as inputs and outputs a score value, which is high when the image-pose pair matches and low otherwise. The network structure consists of a convolutional neural network for image feature extraction, followed by two sub-networks for transforming the image features and pose into a joint embedding. The score function is then the dot-product between the image and pose embeddings. The image-pose embedding and score function are jointly trained using a maximum-margin cost function. Our proposed framework can be interpreted as a special form of structured support vector machines where the joint feature space is discriminatively learned using deep neural networks. We test our framework on the Human3.6m dataset and obtain state-of-the-art results compared to other recent methods. Finally, we present visualizations of the image-pose embedding space, demonstrating the network has learned a high-level embedding of body-orientation and pose-configuration.

****

An Exploration of Parameter Redundancy in Deep Networks With Circulant Projections

Yu Cheng, Felix X. Yu, Rogerio S. Feris, Sanjiv Kumar, Alok Choudhary, Shi-Fu Chang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2857-2865

We explore the redundancy of parameters in deep neural networks by replacing the conventional linear projection in fully-connected layers with the circulant projection. The circulant structure substantially reduces memory footprint and enables the use of the Fast Fourier Transform to speed up the computation. Considering a fully-connected neural network layer with d input nodes, and d output nodes, this method improves the time complexity from O(d^2) to O(dlogd) and space complexity from O(d^2) to O(d). The space savings are particularly important for modern deep convolutional neural network architectures, where fully-connected layers typically contain more than 90% of the network parameters. We further show th at the gradient computation and optimization of the circulant projections can be performed very efficiently. Our experiments on three standard datasets show that the proposed approach achieves this significant gain in storage and efficiency with minimal increase in error rate compared to neural networks with unstructured projections.

****

Additive Nearest Neighbor Feature Maps

Zhenzhen Wang, Xiao-Tong Yuan, Qingshan Liu, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2866-2874

In this paper, we present a concise framework to approximately construct feature maps for nonlinear additive kernels such as the Intersection, Hellinger's, and Chi^2 kernels. The core idea is to construct for each individual feature a set of anchor points and assign to every query the feature map of its nearest neighbor or the weighted combination of those of its k-nearest neighbors in the anchors. The resultant feature maps can be compactly stored by a group of nearest neighbor (binary) indication vectors along with the anchor feature maps. The approximation error of such an anchored feature mapping approach is analyzed. We evaluate the performance of our approach on large-scale nonlinear support vector machines (SVMs) learning tasks in the context of visual object classification. Experimental results on several benchmark data sets show the superiority of our method over existing feature mapping methods in achieving reasonable trade-off between training time and testing accuracy.

****

Understanding Deep Features With Computer-Generated Imagery

Mathieu Aubry, Bryan C. Russell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2875-2883

We introduce an approach for analyzing the variation of features generated by convolutional neural networks (CNNs) trained on large image datasets with respect to scene factors that occur in natural images. Such factors may include object style, 3D viewpoint, color, and scene lighting configuration. Our approach analyzes CNN feature responses with respect to different scene factors by controlling for them via rendering using a large database of 3D CAD models. The rendered images are presented to a trained CNN and responses for different layers are studied with respect to the input scene factors. We perform a linear decomposition of the responses based on knowledge of the input scene factors and analyze the resulting components. In particular, we quantify their relative importance in the CNN responses and visualize them using principal component analysis. We show qualitative and quantitative results of our study on three trained CNNs: AlexNet [??], Places [??], and Oxford VGG [??]. We observe important differences across the different networks and CNN layers with respect to different scene factors and object categories. Finally, we demonstrate that our analysis based on computer-generated imagery translates to the network representation of natural images.

****

Interpolation on the Manifold of K Component GMMs

Hyunwoo J. Kim, Nagesh Adluru, Monami Banerjee, Baba C. Vemuri, Vikas Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2884-2892

Probability density functions (PDFs) are fundamental "objects" in mathematics with numerous applications in computer vision, machine learning and medical imaging. The feasibility of basic operations such as computing the distance between two PDFs and estimating a mean of a set of PDFs is a direct function of the representation we choose to work with. In this paper, we study the Gaussian mixture model (GMM) representation of the PDFs motivated by its numerous attractive features. (1) GMMs are arguably more interpretable than, say, square root parameterizations (2) the model complexity can be explicitly controlled by the number of components and (3) they are already widely used in many applications. The main contributions of this paper are numerical algorithms to enable basic operations on such objects that strictly respect their underlying geometry. For instance, when operating with a set of k component GMMs, a first order expectation is that the result of simple operations like interpolation and averaging should provid e an object that is also a k component GMM. The literature provides very little guidance on enforcing such requirements systematically. It turns out that these tasks are important internal modules for analysis and processing of a field of ensemble average propagators (EAPs), common in diffusion weighted magnetic resonance imaging. We provide proof of principle experiments showing how the proposed algorithms for interpolation can facilitate statistical analysis of such data, essential to many neuroimaging studies. Separately, we also derive interesting connections of our algorithm with functional spaces of Gaussians, that may be of independent interest.

****

Context-Aware CNNs for Person Head Detection

Tuan-Hung Vu, Anton Osokin, Ivan Laptev; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2893-2901

Person detection is a key problem for many computer vision tasks. While face detection has reached maturity, detecting people under full variation of camera view-points, human poses, lighting conditions and occlusions is still a difficult challenge. In this work we focus on detecting human heads in natural scenes. Starting from the recent R-CNN object detector, we extend it in two ways. First, we leverage person-scene relations and propose a global CNN model trained to predict positions and scales of heads directly from the full image. Second, we explicitly model pairwise relations among the objects via energy-based model where the potentials are computed with a CNN framework. Our full combined model complements R-CNN with contextual cues derived from the scene. To train and test our model, we introduce a large dataset with 369,846 human heads annotated in 224,740 movie frames. We evaluate our method and demonstrate improvements of person head detection compared to several recent baselines on three datasets. We also show improvements of the detection speed provided by our model.

****

Mode-Seeking on Hypergraphs for Robust Geometric Model Fitting

Hanzi Wang, Guobao Xiao, Yan Yan, David Suter; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2902-2910

In this paper, we propose a novel geometric model fitting method, called Mode-Seeking on Hypergraphs (MSH), to deal with multi-structure data even in the presence of severe outliers. The proposed method formulates geometric model fitting as a mode seeking problem on a hypergraph in which vertices represent model hypotheses and hyperedges denote data points. MSH intuitively detects model instances by a simple and effective mode seeking algorithm. In addition to the mode seeking algorithm, MSH includes a similarity measure between vertices on the hypergraph and a "weight-aware sampling" technique. The proposed method not only alleviates sensitivity to the data distribution, but also is scalable to large scale problems. Experimental results further demonstrate that the proposed method has significant superiority over the state-of-the-art fitting methods on both synthetic data and real images.

****

Highly-Expressive Spaces of Well-Behaved Transformations: Keeping It Simple

Oren Freifeld, Soren Hauberg, Kayhan Batmanghelich, John W. Fisher III; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2911-2919

We propose novel finite-dimensional spaces of R - R transformations, n [?] 1, 2, 3, derived from (continuously-defined) parametric stationary velocity fields. Particularly, we obtain these transformations, which are diffeomorphisms, by fast and highly-accurate integration of continuous piecewise-affine velocity fields; we also provide an exact solution for n = 1. The simple-yet-highly-expressive proposed representation handles optional constraints (e.g., volume preservation) easily and supports convenient modeling choices and rapid likelihood evaluations (facilitating tractable inference over latent transformations). Its applications include, but are not limited to: unconstrained optimization over monotonic functions; modeling cumulative distribution functions or histograms; time warping; image registration; landmark-based warping; real-time diffeomorphic image editing. Our code is available at https://github.com/freifeld/cpabDiffeo

****

Entropy-Based Latent Structured Output Prediction

Diane Bouchacourt, Sebastian Nowozin, M. Pawan Kumar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2920-2928

Recently several generalizations of the popular latent structural SVM framework have been proposed in the literature. Broadly speaking, the generalizations can be divided into two categories: (i) those that predict the output variables while either marginalizing the latent variables or estimating their most likely values; and (ii) those that predict the output variables by minimizing an entropy-based uncertainty measure over the latent space. In order to aid their application in computer vision, we study these generalizations with the aim of identifying their strengths and weaknesses. To this end, we propose a novel prediction criterion that includes as special cases all previous prediction criteria that have been used in the literature. Specifically, our framework's prediction criterion minimizes the Aczel and Daroczy entropy of the output. This in turn allows us to design a learning objective that provides a unified framework (UF) for latent structured prediction. We develop a single optimization algorithm and empirically show that it is as effective as the more complex approaches that have been previously employed for latent structured prediction. Using this algorithm, we provide empirical evidence that lends support to prediction via the minimization of the latent space uncertainty.

****

Fast Orthogonal Projection Based on Kronecker Product

Xu Zhang, Felix X. Yu, Ruiqi Guo, Sanjiv Kumar, Shengjin Wang, Shi-Fu Chang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2929-2937

We propose a family of structured matrices to speed up orthogonal projections for high-dimensional data commonly seen in computer vision applications. In this, a structured matrix is formed by the Kronecker product of a series of smaller orthogonal matrices. This achieves O(dlogd) computational complexity and O(logd) space complexity for d-dimensional data, a drastic improvement over the standard unstructured projections whose computational and space complexities are both O(d^2). The proposed structured matrices are applicable to a number of application domains, and are faster and more compact than other structured matrices used in the past. We also introduce an efficient learning procedure for optimizing such matrices in a data dependent fashion. We demonstrate the significant advantages of the proposed approach in solving the approximate nearest neighbor (ANN) image search problem with both binary embedding and quantization. We find that the orthogonality plays a very important role in solving ANN problem, since the random orthogonal Kronecker projection has already provided promising performance. Comprehensive experiments show that the proposed approach can achieve similar or better accuracy as the existing state-of-the-art but with significantly less time and memory.

****

PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization

Alex Kendall, Matthew Grimes, Roberto Cipolla; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2938-2946

We present a robust and real-time monocular six degree of freedom relocalization system. Our system trains a convolutional neural network to regress the 6-DOF camera pose from a single RGB image in an end-to-end manner with no need of additional engineering or graph optimisation. The algorithm can operate indoors and outdoors in real time, taking 5ms per frame to compute. It obtains approximately 2m and 3 degrees accuracy for large scale outdoor scenes and 0.5m and 5 degrees accuracy indoors. This is achieved using an efficient 23 layer deep convnet, demonstrating that convnets can be used to solve complicated out of image plane regression problems. This was made possible by leveraging transfer learning from large scale classification data. We show that the PoseNet localizes from high level features and is robust to difficult lighting, motion blur and different camera intrinsics where point based SIFT registration fails. Furthermore we show how the pose feature that is produced generalizes to other scenes allowing us to regr

ess pose with only a few dozen training examples.
****

Predicting Multiple Structured Visual Interpretations
Debadeepta Dey, Varun Ramakrishna, Martial Hebert, J. Andrew Bagnell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2947-2955

We present a simple approach for producing a small number of structured visual outputs which have high recall, for a variety of tasks including monocular pose estimation and semantic scene segmentation. Current state-of-the-art approaches learn a single model and modify inference procedures to produce a small number of diverse predictions. We take the alternate route of modifying the learning procedure to directly optimize for good, high recall sequences of structured-output predictors. Our approach introduces no new parameters, naturally learns diverse predictions and is not tied to any specific structured learning or inference procedure. We leverage recent advances in the contextual submodular maximization literature to learn a sequence of predictors and empirically demonstrate the simplicity and performance of our approach on multiple challenging vision tasks including achieving state-of-the-art results on multiple predictions for monocular pose-estimation and image foreground/background segmentation.
****

Look and Think Twice: Capturing Top-Down Visual Attention With Feedback Convolutional Neural Networks
Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, Thomas S. Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2956-2964

While feedforward deep convolutional neural networks (CNNs) have been a great success in computer vision, it is important to remember that the human visual contex contains generally more feedback connections than foward connections. In this paper, we will briefly introduce the background of feedbacks in the human visual cortex, which motivates us to develop a computational feedback mechanism in the deep neural networks. The proposed networks perform inference from image features in a bottom-up manner as traditional convolutional networks; while during feedback loops it sets up high-level semantic labels as the agoala to infer the activation status of hidden layer neurons. The feedback networks help us better visualize and understand on how deep neural networks work as well as capture visual attention on expected objects, even in the images with cluttered background and multiple objects.
****

Matrix Backpropagation for Deep Networks With Structured Layers
Catalin Ionescu, Orestis Vantzos, Cristian Sminchisescu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2965-2973

Deep neural network architectures have recently produced excellent results in a variety of areas in artificial intelligence and visual recognition, well surpassing traditional shallow architectures trained using hand-designed features. The power of deep networks stems both from their ability to perform local computations followed by pointwise non-linearities over increasingly larger receptive fields, and from the simplicity and scalability of the gradient-descent training procedure based on backpropagation. An open problem is the inclusion of layers that perform global, structured matrix computations like segmentation (e.g. normalized cuts) or higher-order pooling (e.g. log-tangent space metrics defined over the manifold of symmetric positive definite matrices) while preserving the validity and efficiency of an end-to-end deep training framework. In this paper we propose a sound mathematical apparatus to formally integrate global structured computation into deep computation architectures. At the heart of our methodology is the development of the theory and practice of backpropagation that generalizes to the calculus of adjoint matrix variations. We perform segmentation experiments using the BSDS and MSCOCO benchmarks and demonstrate that deep networks relying on second-order pooling and normalized cuts layers, trained end-to-end using matrix backpropagation, outperform counterparts that do not take advantage of suc h global layers.

****

Introducing Geometry in Active Learning for Image Segmentation

Ksenia Konyushkova, Raphael Sznitman, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2974-2982

We propose an Active Learning approach to training a segmentation classifier that exploits geometric priors to streamline the annotation process in 3D image volumes. To this end, we use these priors not only to select voxels most in need of annotation but to guarantee that they lie on 2D planar patch, which makes it much easier to annotate than if they were randomly distributed in the volume. A simplified version of this approach is effective in natural 2D images. We evaluated our approach on Electron Microscopy and Magnetic Resonance image volumes, as well as on natural images. Comparing our approach against several accepted baselines demonstrates a marked performance increase.

****

Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition

Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, Junmo Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983-2991

Temporal information has useful features for recognizing facial expressions. However, to manually design useful features requires a lot of effort. In this paper, to reduce this effort, a deep learning technique, which is regarded as a tool to automatically extract useful features from raw data, is adopted. Our deep network is based on two different models. The first deep network extracts temporal appearance features from image sequences, while the other deep network extracts temporal geometry features from temporal facial landmark points. These two models are combined using a new integration method in order to boost the performance of the facial expression recognition. Through several experiments, we show that the two models cooperate with each other. As a result, we achieve superior performance to other state-of-the-art methods in the CK+ and Oulu-CASIA databases. Furthermore, we show that our new integration method gives more accurate results than traditional methods, such as a weighted summation and a feature concatenation method.

****

Direct Intrinsics: Learning Albedo-Shading Decomposition by Convolutional Regression

Takuya Narihira, Michael Maire, Stella X. Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2992-2992

We introduce a new approach to intrinsic image decomposition, the task of decomposing a single image into albedo and shading components. Our strategy, which we term direct intrinsics, is to learn a convolutional neural network (CNN) that directly predicts output albedo and shading channels from an input RGB image patch. Direct intrinsics is a departure from classical techniques for intrinsic image decomposition, which typically rely on physically-motivated priors and graph-based inference algorithms. The large-scale synthetic ground-truth of the MPI Sintel dataset plays the key role in training direct intrinsics. We demonstrate results on both the synthetic images of Sintel and the real images of the classic MIT intrinsic image dataset. On Sintel, direct intrinsics, using only RGB input, outperforms all prior work, including methods that rely on RGB+Depth input. Direct intrinsics also generalizes across modalities; our Sintel-trained CNN produces quite reasonable decompositions on the real images of the MIT dataset. Our results indicate that the marriage of CNNs with synthetic training data may be a powerful new technique for tackling classic problems in computer vision.

****

Face Flow

Patrick Snape, Anastasios Roussos, Yannis Panagakis, Stefanos Zafeiriou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2993-3001

In this paper, we propose a method for the robust and efficient computation of multi-frame optical flow in an expressive sequence of facial images. We formulate a novel energy minimisation problem for establishing dense correspondences be tween a neutral template and every frame of a sequence. We exploit the highly correlated nature of human expressions by representing dense facial motion using a deformation basis. Furthermore, we exploit the even higher correlation between deformations in a given input sequence by imposing a low-rank prior on the coefficients of the deformation basis, yielding temporally consistent optical flow. Our proposed model-based formulation, in conjunction with the inverse compositional strategy and low-rank matrix optimisation that we adopt, leads to a highly efficient algorithm for calculating facial flow. As experimental evaluation, we show quantitative experiments on a challenging novel benchmark of face sequences, with dense ground truth optical flow provided by motion capture data. We also provide qualitative results on a real sequence displaying fast motion and occlusions. Extensive quantitative and qualitative comparisons demonstrate that the proposed method outperforms state-of-the-art optical flow and dense non-rigid registration techniques, whilst running an order of magnitude faster.

****

Discriminative Low-Rank Tracking

Yao Sui, Yafei Tang, Li Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3002-3010

Good tracking performance is in general attributed to accurate representation over previously obtained targets or reliable discrimination between the target and the surrounding background. In this work, we exploit the advantages of the both approaches to achieve a robust tracker. We construct a subspace to represent the target and the neighboring background, and simultaneously propagate their class labels via the learned subspace. Moreover, we propose a novel criterion to identify the target from numerous target candidates on each frame, which takes into account both discrimination reliability and representation accuracy. In addition, with the proposed criterion, the ambiguity in the class labels of the neighboring background samples, which often influences the reliability of discriminative tracking model, is effectively alleviated, while the training set is still kept small. Extensive experiments demonstrate that our tracker performs favourably against many other state-of-the-art trackers.

****

SOWP: Spatially Ordered and Weighted Patch Descriptor for Visual Tracking

Han-Ul Kim, Dae-Youn Lee, Jae-Young Sim, Chang-Su Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3011-3019

A simple yet effective object descriptor for visual tracking is proposed in this paper. We first decompose the bounding box of a target object into multiple patches, which are described by color and gradient histograms. Then, we concatenate the features of the spatially ordered patches to represent the object appearance. Moreover, to alleviate the impacts of background information possibly included in the bounding box, we determine patch weights using random walk with restart (RWR) simulations. The patch weights represent the importance of each patch in the description of foreground information, and are used to construct an object descriptor, called spatially ordered and weighted patch (SOWP) descriptor. We incorporate the proposed SOWP descriptor into the structured output tracking framework. Experimental results demonstrate that the proposed algorithm yields significantly better performance than the state-of-the-art trackers on a recent benchmark dataset, and also excels in another recent benchmark dataset.

****

Live Repetition Counting

Ofir Levy, Lior Wolf; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3020-3028

The task of counting the number of repetitions of approximately the same action in an input video sequence is addressed. The proposed method runs online and not on the complete pre-captured video. It analyzes sequentially blocks of 20 non-consecutive frames. The cycle length within each block is evaluated using a convolutional neural network and the information is then integrated over time. The entropy of the network's predictions is used in order to automatically start and stop the repetition counter and to select the appropriate time scale. Coupled with a region of interest detection mechanism, the method is robust enough to handl e real world videos, even when the camera is moving. A unique property of our method is that it is shown to successfully train on entirely unrealistic data created by synthesizing moving random patches.

****

Near-Online Multi-Target Tracking With Aggregated Local Flow Descriptor

Wongun Choi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3029-3037

In this paper, we tackle two key aspects of multiple target tracking problem: 1) designing an accurate affinity measure to associate detections and 2) implementing an efficient and accurate (near) online multiple target tracking algorithm. As for the first contribution, we introduce a novel Aggregated Local Flow Descriptor (ALFD) that encodes the relative motion pattern between a pair of temporally distant detections using long term interest point trajectories (IPTs). Leveraging on the IPTs, the ALFD provides a robust affinity measure for estimating the likelihood of matching detections regardless of the application scenarios. As for another contribution, we present a Near-Online Multi-target Tracking (NOMT) algorithm. The tracking problem is formulated as a data-association between targets and detections in a temporal window, that is performed repeatedly at every frame. While being efficient, NOMT achieves robustness via integrating multiple cues including ALFD metric, target dynamics, appearance similarity, and long term trajectory regularization into the model. Our ablative analysis verifies the superiority of the ALFD metric over the other conventional affinity metrics. We run a comprehensive experimental evaluation on two challenging tracking datasets, KITTI and MOT datasets. The NOMT method combined with ALFD metric achieves the best accuracy in both datasets with significant margins (about 10% higher MOTA) over the state-of-the-art.

****

Multi-Kernel Correlation Filter for Visual Tracking

Ming Tang, Jiayi Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3038-3046

Correlation filter based trackers are ranked top in terms of performances. Nevertheless, they only employ a single kernel at a time. In this paper, we will derive a multi-kernel correlation filter (MKCF) based tracker which fully takes advantage of the invariance-discriminative power spectrums of various features to further improve the performance. Moreover, it may easily introduce location and representation errors to search several discrete scales for the proper one of the object bounding box, because normally the discrete candidate scales are determined and the corresponding feature pyramid are generated ahead of searching. In this paper, we will propose a novel and efficient scale estimation method based on optimal bisection search and fast evaluation of features. Our scale estimation method is the first one that uses the truly minimal number of layers of feature pyramid and avoids constructing the pyramid before searching for proper scales.

****

Joint Probabilistic Data Association Revisited

Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, Ian Reid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3047-3055

In this paper, we revisit the joint probabilistic data association (JPDA) technique and propose a novel solution based on recent developments in finding the m-best solutions to an integer linear program. The key advantage of this approach is that it makes JPDA computationally tractable in applications with high target and/or clutter density, such as spot tracking in fluorescence microscopy sequences and pedestrian tracking in surveillance footage. We also show that our JPDA algorithm embedded in a simple tracking framework is surprisingly competitive with state-of-the-art global tracking methods in these two applications, while needing considerably less processing time.

****

Tracking-by-Segmentation With Online Gradient Boosting Decision Tree

Jeany Son, Ilchae Jung, Kayoung Park, Bohyung Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3056-3064

We propose an online tracking algorithm that adaptively models target appearances based on an online gradient boosting decision tree. Our algorithm is particularly useful for non-rigid and/or articulated objects since it handles various deformations of the target effectively by integrating a classifier operating on individual patches and provides segmentation masks of the target as final results. The posterior of the target state is propagated over time by particle filtering, where the likelihood is computed based mainly on patch-level confidence map associated with a latent target state corresponding to each sample. Once tracking is completed in each frame, our gradient boosting decision tree is updated to adapt new data in a recursive manner. For effective evaluation of segmentation-based tracking algorithms, we construct a new ground-truth that contains pixel-level annotation of segmentation mask. We evaluate the performance of our tracking algorithm based on the measures for segmentation masks, where our algorithm illustrates superior accuracy compared to the state-of-the-art segmentation-based tracking methods.

****

Exploring Causal Relationships in Visual Object Tracking

Karel Lebeda, Simon Hadfield, Richard Bowden; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3065-3073

Causal relationships can often be found in visual object tracking between the motions of the camera and that of the tracked object. This object motion may be an effect of the camera motion, e.g. an unsteady handheld camera. But it may also be the cause, e.g. the cameraman framing the ob- ject. In this paper we explore these relationships, and pro- vide statistical tools to detect and quantify them; these are based on transfer entropy and stem from information the- ory. The relationships are then exploited to make predic- tions about the object location. The approach is shown to be an excellent measure for describing such relationships. On the VOT2013 dataset the prediction accuracy is increased by 62 % over the best non-causal predictor. We show that the location predictions are robust to camera shake and sud- den motion, which is invaluable for any tracking algorithm and demonstrate this by applying causal prediction to two state-of-the-art trackers. Both of them benefit, Struck gain- ing a 7 % accuracy and 22 % robustness increase on the VTB1.1 benchmark, becoming the new state-of-the-art.

****

Hierarchical Convolutional Features for Visual Tracking

Chao Ma, Jia-Bin Huang, Xiaokang Yang, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3074-3082

Visual object tracking is challenging as target objects often undergo significant appearance changes caused by deformation, abrupt motion, background clutter and occlusion. In this paper, we exploit features extracted from deep convolutional neural networks trained on object recognition datasets to improve tracking accuracy and robustness. The outputs of the last convolutional layers encode the semantic information of targets and such representations are robust to significant appearance variations. However, their spatial resolution is too coarse to precisely localize targets. In contrast, earlier convolutional layers provide more precise localization but are less invariant to appearance changes. We interpret the hierarchies of convolutional layers as a nonlinear counterpart of an image pyramid representation and exploit these multiple levels of abstraction for visual tracking. Specifically, we adaptively learn correlation filters on each convolutional layer to encode the target appearance. We hierarchically infer the maximum response of each layer to locate targets. Extensive experimental results on a largescale benchmark dataset show that the proposed algorithm performs favorably against state-of-the-art methods.

****

Robust Non-Rigid Motion Tracking and Surface Reconstruction Using L0 Regularization

Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, Qionghai Dai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3083-3091

We present a new motion tracking method to robustly reconstruct non-rigid geometries and motions from single view depth inputs captured by a consumer depth sens or. The idea comes from the observation of the existence of intrinsic articulated subspace in most of non-rigid motions. To take advantage of this characteristic, we propose a novel L0 based motion regularizer with an iterative optimization solver that can implicitly constrain local deformation only on joints with articulated motions, leading to reduced solution space and physical plausible deformations. The L0 strategy is integrated into the available non-rigid motion tracking pipeline, forming the proposed L0-L2 non-rigid motion tracking method that can adaptively stop the tracking error propagation. Extensive experiments over complex human body motions with occlusions, face and hand motions demonstrate that our approach substantially improves tracking robustness and surface reconstruction accuracy.

****

Online Object Tracking With Proposal Selection
Yang Hua, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3092-3100
Tracking-by-detection approaches are some of the most successful object trackers in recent years. Their success is largely determined by the detector model they learn initially and then update over time. However, under challenging conditions where an object can undergo transformations, e.g., severe rotation, these methods are found to be lacking. In this paper, we address this problem by formulating it as a proposal selection task and making two contributions. The first one is introducing novel proposals estimated from the geometric transformations undergone by the object, and building a rich candidate set for predicting the object location. The second one is devising a novel selection strategy using multiple cues, i.e., detection score and edgeness score computed from state-of-the-art object edges and motion boundaries. We extensively evaluate our approach on the visual object tracking 2014 challenge and online tracking benchmark datasets, and show the best performance.

****

Understanding and Diagnosing Visual Tracking Systems
Naiyan Wang, Jianping Shi, Dit-Yan Yeung, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3101-3109
Several benchmark datasets for visual tracking research have been created in recent years. Despite their usefulness, whether they are sufficient for understanding and diagnosing the strengths and weaknesses of different trackers remains questionable. To address this issue, we propose a framework by breaking a tracker down into five constituent parts, namely, motion model, feature extractor, observation model, model updater, and ensemble post-processor. We then conduct ablative experiments on each component to study how it affects the overall result. Surprisingly, our findings are discrepant with some common beliefs in the visual tracking research community. We find that the feature extractor plays the most important role in a tracker. On the other hand, although the observation model is the focus of many studies, we find that it often brings no significant improvement. Moreover, the motion model and model updater contain many details that could affect the result. Also, the ensemble post-processor can improve the result substantially when the constituent trackers have high diversity. Based on our findings, we put together some very elementary building blocks to give a basic tracker which is competitive in performance to the state-of-the-art trackers. We believe our framework can provide a solid baseline when conducting controlled experiments for visual tracking research.

****

Integrating Dashcam Views Through Inter-Video Mapping
Hsin-I Chen, Yi-Ling Chen, Wei-Tse Lee, Fan Wang, Bing-Yu Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3110-3118
In this paper, an inter-video mapping approach is proposed to integrate video footages from two dashcams installed on a preceding and its following vehicle to provide the illusion that the driver of the following vehicle can see-through the preceding one. The key challenge is to adapt the perspectives of the two videos based on a small number of common features since a large portion of the common region in the video captured by the following vehicle is occluded by the precedi

ng one. Inspired by the observation that images with the most similar viewpoints yield dense and high-quality matches, the proposed inter-video mapping estimates spatially-varying motions across two videos utilizing images of very similar contents. Specifically, we estimate frame-to-frame motions of each two consecutive images and incrementally add new views into a merged representation. In this way, long-rang motion estimation is achieved, and the observed perspective discrepancy between the two videos can be well approximated our motion estimation. Once the inter-video mapping is established, the correspondences can be updated incrementally, so the proposed method is suitable for on-line applications. Our experiments demonstrate the effectiveness of our approach on real-world challenging videos.

****

Visual Tracking With Fully Convolutional Networks

Lijun Wang, Wanli Ouyang, Xiaogang Wang, Huchuan Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3119-3127

We propose a new approach for general object tracking with fully convolutional neural network. Instead of treating convolutional neural network (CNN) as a black-box feature extractor, we conduct in-depth study on the properties of CNN features offline pre-trained on massive image data and classification task on ImageNet. The discoveries motivate the design of our tracking system. It is found that convolutional layers in different levels characterize the target from different perspectives. A top layer encodes more semantic features and serves as a category detector, while a lower layer carries more discriminative information and can better separate the target from distracters with similar appearance. Both layers are jointly used with a switch mechanism during tracking. It is also found that for a tracking target, only a subset of neurons are relevant. A feature map selection method is developed to remove noisy and irrelevant feature maps, which can reduce computation redundancy and improve tracking accuracy. Extensive evaluation on the widely used tracking benchmark shows that the proposed tacker outperforms the state-of-the-art significantly.

****

Multiple Feature Fusion via Weighted Entropy for Visual Tracking

Lin Ma, Jiwen Lu, Jianjiang Feng, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3128-3136

It is desirable to combine multiple feature descriptors to improve the visual tracking performance because different features can provide complementary information to describe objects of interest. However, how to effectively fuse multiple features remains a challenging problem in visual tracking, especially in a data-driven manner. In this paper, we propose a new data-adaptive visual tracking approach by using multiple feature fusion via weighted entropy. Unlike existing visual trackers which simply concatenate multiple feature vectors together for object representation, we employ the weighted entropy to evaluate the dissimilarity between the object state and the background state, and seek the optimal feature combination by minimizing the weighted entropy, so that more complementary information can be exploited for object representation. Experimental results demonstrate the effectiveness of our approach in tackling various challenges for visual object tracking.

****

Pedestrian Travel Time Estimation in Crowded Scenes

Shuai Yi, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3137-3145

In this paper, we target on the problem of estimating the statistic of pedestrian travel time within a period from an entrance to a destination in a crowded scene. Such estimation is based on the global distributions of crowd densities and velocities instead of complete trajectories of pedestrians, which cannot be obtained in crowded scenes. The proposed method is motivated by our statistical investigation into the correlations between travel time and global properties of crowded scenes. Active regions are created for each source-destination pair to model the probable walking regions over the corresponding source-destination traffic flow. Two sets of scene features are specially designed for modeling moving and stationary persons inside the active regions and their influences on pedestrian travel time. The estimation of pedestrian travel time provides valuable information for both crowd scene understanding and pedestrian behavior analysis, but was not sufficiently studied in literature. The effectiveness of the proposed pedestrian travel time estimation model is demonstrated through several surveillance applications, including dynamic scene monitoring, localization of regions blocking traffics, and detection of abnormal pedestrian behaviors. Many more valuable applications based on our method are to be explored in the future.

****

Unsupervised Synchrony Discovery in Human Interaction

Wen-Sheng Chu, Jiabei Zeng, Fernando De la Torre, Jeffrey F. Cohn, Daniel S. Messinger; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3146-3154

People are inherently social. Social interaction plays an important and natural role in human behavior. Most computational methods focus on individuals alone rather than in social context. They also require labelled training data. We present an unsupervised approach to discover interpersonal synchrony, referred as to two or more persons preforming common actions in overlapping video frames or segments. For computational efficiency, we develop a branch-and-bound (B&B) approach that affords exhaustive search while guaranteeing a globally optimal solution. The proposed method is entirely general. It takes from two or more videos any multi-dimensional signal that can be represented as a histogram. We derive three novel bounding functions and provide efficient extensions, including multi-synchrony detection and accelerated search, using a warm-start strategy and parallelism. We evaluate the effectiveness of our approach in multiple databases, including human actions using the CMU Mocap dataset, spontaneous facial behaviors using group-formation task dataset and parent-infant interaction dataset.

****

Efficient Video Segmentation Using Parametric Graph Partitioning

Chen-Ping Yu, Hieu Le, Gregory Zelinsky, Dimitris Samaras; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3155-3163

Video segmentation is the task of grouping similar pixels in the spatio-temporal domain, and has become an important preprocessing step for subsequent video analysis. Most video segmentation and supervoxel methods output a hierarchy of segmentations, but while this provides useful multiscale information, it also adds difficulty in selecting the appropriate level for a task. In this work, we propose an efficient and robust video segmentation framework based on parametric graph partitioning (PGP), a fast, almost parameter free graph partitioning method that identifies and removes between-cluster edges to form node clusters. Apart from its computational efficiency, PGP performs clustering of the spatio-temporal volume without requiring a pre-specified cluster number or bandwidth parameters, thus making video segmentation more practical to use in applications. The PGP framework also allows processing sub-volumes, which further improves performance, contrary to other streaming video segmentation methods where sub-volume processing reduces performance. We evaluate the PGP method using the SegTrack v2 and Chen Xiph.org datasets, and show that it outperforms related state-of-the-art algorithms in 3D segmentation metrics and running time.

****

Learning to Track for Spatio-Temporal Action Localization

Philippe Weinzaepfel, Zaid Harchaoui, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3164-3172

We propose an effective approach for spatio-temporal action localization in realistic videos. The approach first detects proposals at the frame-level and scores them with a combination of static and motion CNN features. It then tracks high-scoring proposals throughout the video using a tracking-by-detection approach. Our tracker relies simultaneously on instance-level and class-level detectors. The tracks are scored using a spatio-temporal motion histogram, a descriptor at the track level, in combination with the CNN features. Finally, we perform temporal localization of the action using a sliding-window approach at the track level. We present experimental results for spatio-temporal localization on the UCF-Sp orts, J-HMDB and UCF-101 action localization datasets, where our approach outperforms the state of the art with a margin of 15%, 7% and 12% respectively in mAP.

****

Unsupervised Object Discovery and Tracking in Video Collections

Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3173-3181

This paper addresses the problem of automatically localizing dominant objects as spatio-temporal tubes in a noisy collection of videos with minimal or even no supervision. We formulate the problem as a combination of two complementary processes: discovery and tracking. The first one establishes correspondences between prominent regions across videos, and the second one associates similar object regions within the same video. Interestingly, our algorithm also discovers the implicit topology of frames associated with instances of the same object class across different videos, a role normally left to supervisory information in the form of class labels in conventional image and video understanding methods. Indeed, as demonstrated by our experiments, our method can handle video collections featuring multiple object classes, and substantially outperforms the state of the art in colocalization, even though it tackles a broader problem with much less supervision.

****

Car That Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models

Ashesh Jain, Hema S. Koppula, Bharad Raghavan, Shane Soh, Ashutosh Saxena; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3182-3190

Advanced Driver Assistance Systems (ADAS) have made driving safer over the last decade. They prepare vehicles for unsafe road conditions and alert drivers if they perform a dangerous maneuver. However, many accidents are unavoidable because by the time drivers are alerted, it is already too late. Anticipating maneuvers beforehand can alert drivers before they perform the maneuver and also give ADAS more time to avoid or prepare for the danger. In this work we anticipate driving maneuvers a few seconds before they occur. For this purpose we equip a car with cameras and a computing device to capture the driving context from both inside and outside of the car. We propose an Autoregressive Input-Output HMM to model the contextual information alongwith the maneuvers. We evaluate our approach on a diverse data set with 1180 miles of natural freeway and city driving and show that we can anticipate maneuvers 3.5 seconds before they occur with over 80% F1-score in real-time.

****

Activity Auto-Completion: Predicting Human Activities From Partial Videos

Zhen Xu, Laiyun Qing, Jun Miao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3191-3199

In this paper, we propose an activity auto-completion (AAC) model for human activity prediction by formulating activity prediction as a query auto-completion (QAC) problem in information retrieval. First, we extract discriminative patches in frames of videos. A video is represented based on these patches and divided into a collection of segments, each of which is regarded as a character typed in the search box. Then a partially observed video is considered as an activity prefix, consisting of one or more characters. Finally, the missing observation of an activity is predicted as the activity candidates provided by the auto-completion model. The candidates are matched against the activity prefix on-the-fly and ranked by a learning-to-rank algorithm. We validate our method on UT-Interaction Set #1 and Set #2 [19]. The experimental results show that the proposed activity auto-completion model achieves promising performance.

****

Person Re-Identification With Correspondence Structure Learning

Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, Jingdong Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3200-3208

This paper addresses the problem of handling spatial misalignments due to camera

-view changes or human-pose variations in person re-identification. We first introduce a boosting-based approach to learn a correspondence structure which indicates the patch-wise matching probabilities between images from a target camera pair. The learned correspondence structure can not only capture the spatial correspondence pattern between cameras but also handle the viewpoint or human-pose variation in individual images. We further introduce a global-based matching process. It integrates a global matching constraint over the learned correspondence structure to exclude cross-view misalignments during the image patch matching process, hence achieving a more reliable matching score between images. Experimental results on various datasets demonstrate the effectiveness of our approach.
****

Adaptive Exponential Smoothing for Online Filtering of Pixel Prediction Maps
Kang Dang, Jiong Yang, Junsong Yuan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3209-3217

We propose an efficient online video filtering method, called adaptive exponential filtering (AES) to refine pixel prediction maps. Assuming each pixel is associated with a discriminative prediction score, the proposed AES applies exponentially decreasing weights over time to smooth the prediction score of each pixel, similar to classic exponential smoothing. However, instead of fixing the spatial pixel location to perform temporal filtering, we trace each pixel in the past frames by finding the optimal path that can bring the maximum exponential smoothing score, thus performing adaptive and non-linear filtering. Thanks to the pixel tracing, AES can better address object movements and avoid over-smoothing. To enable real-time filtering, we propose a linear-complexity dynamic programming scheme that can trace all pixels simultaneously. We apply the proposed filtering method to improve both saliency detection maps and scene parsing maps. The comparisons with average and exponential filtering, as well as state-of-the-art methods, validate that our AES can effectively refine the pixel prediction maps, without using the original video again.
****

P-CNN: Pose-Based CNN Features for Action Recognition
Guilhem Cheron, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3218-3226

This work targets human action recognition in video. While recent methods typically represent actions by statistics of local video features, here we argue for the importance of a representation derived from human pose. To this end we propose a new Pose-based Convolutional Neural Network descriptor (P-CNN) for action recognition. The descriptor aggregates motion and appearance information along tracks of human body parts. We investigate different schemes of temporal aggregation and experiment with P-CNN features obtained both for automatically estimated and manually annotated human poses. We evaluate our method on the recent and challenging JHMDB and MPII Cooking datasets. For both datasets our method shows consistent improvement over the state of the art.
****

Fully Connected Object Proposals for Video Segmentation
Federico Perazzi, Oliver Wang, Markus Gross, Alexander Sorkine-Hornung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3227-3234

We present a novel approach to video segmentation using multiple object proposals. The problem is formulated as a minimization of a novel energy function defined over a fully connected graph of object proposals. Our model combines appearance with long-range point tracks, which is key to ensure robustness with respect to fast motion and occlusions over longer video sequences. As opposed to previous approaches based on object proposals, we do not seek the best per-frame object hypotheses to perform the segmentation. Instead, we combine multiple, potentially imperfect proposals to improve overall segmentation accuracy and ensure robustness to outliers. Overall, the basic algorithm consists of three steps. First, we generate a very large number of object proposals for each video frame using existing techniques. Next, we perform an SVM-based pruning step to retain only high quality proposals with sufficiently discriminative power. Finally, we determin e the fore- and background classification by solving for the maximum a posteriori of a fully connected conditional random field, defined using our novel energy function. Experimental results on a well established dataset demonstrate that our method compares favorably to several recent state-of-the-art approaches.

****

Video Segmentation With Just a Few Strokes

Naveen Shankar Nagaraja, Frank R. Schmidt, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3235-3243

As the use of videos is becoming more popular in computer vision, the need for annotated video datasets increases. Such datasets are required either as training data or simply as ground truth for benchmark datasets. A particular challenge in video segmentation is due to disocclusions, which hamper frame-to-frame propagation, in conjunction with non-moving objects. We show that a combination of motion from point trajectories, as known from motion segmentation, along with minimal supervision can largely help solve this problem. Moreover, we integrate a new constraint that enforces consistency of the color distribution in successive frames. We quantify user interaction effort with respect to segmentation quality on challenging ego motion videos. We compare our approach to a diverse set of algorithms in terms of user effort and in terms of performance on common video segmentation benchmarks.

****

Actionness-Assisted Recognition of Actions

Ye Luo, Loong-Fah Cheong, An Tran; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3244-3252

We elicit from a fundamental definition of action low-level attributes that can reveal agency and intentionality. These descriptors are mainly trajectory-based, measuring sudden changes, temporal synchrony, and repetitiveness. The actionness map can be used to localize actions in a way that is generic across action and agent types. Furthermore, it also groups interacting regions into a useful unit of analysis, which is crucial for recognition of actions involving interactions. We then implement an actionness-driven pooling scheme to improve action recognition performance. Experimental results on three datasets show the advantages of our method on both action detection and action recognition comparing with other state-of-the-art methods.

****

COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation

Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, Ryuzo Okada; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3253-3261

This paper presents a patch-based approach for crowd density estimation in public scenes. We formulate the problem of estimating density in a structured learning framework applied to random decision forests. Our approach learns the mapping between patch features and relative locations of all objects inside each patch, which contribute to generate the patch density map through Gaussian kernel density estimation. We build the forest in a coarse-to-fine manner with two split node layers, and further propose a crowdedness prior and an effective forest reduction method to improve the estimation accuracy and speed. Moreover, we introduce a semi-automatic training method to learn the estimator for a specific scene. We achieved state-of-the-art results on the public Mall dataset and UCSD dataset, and also proposed two potential applications in traffic counts and scene understanding with promising results.

****

Multi-Cue Structure Preserving MRF for Unconstrained Video Segmentation

Saehoon Yi, Vladimir Pavlovic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3262-3270

Video segmentation is a stepping stone to understanding video context. Video segmentation enables one to represent a video by decomposing it into coherent regions which comprise whole or parts of objects. However, the challenge originates from the fact that most of the video segmentation algorithms are based on unsupervised learning due to expensive cost of pixelwise video annotation and intra-c lass variability within similar unconstrained video classes. We propose a Markov Random Field model for unconstrained video segmentation that relies on tight integration of multiple cues: vertices are defined from contour based superpixels, unary potentials from temporally smooth label likelihood and pairwise potentials from global structure of a video. Multi-cue structure is a breakthrough to extracting coherent object regions for unconstrained videos in absence of supervision. Our experiments on VSB100 dataset show that the proposed model significantly outperforms competing state-of-the-art algorithms. Qualitative analysis illustrates that video segmentation result of the proposed model is consistent with human perception of objects.

****

Motion Trajectory Segmentation via Minimum Cost Multicuts

Margret Keuper, Bjoern Andres, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3271-3279

For the segmentation of moving objects in videos, the analysis of long-term point trajectories has been very popular recently. In this paper, we formulate the segmentation of a video sequence based on point trajectories as a minimum cost multicut problem. Unlike the commonly used spectral clustering formulation, the minimum cost multicut formulation gives natural rise to optimize not only for a cluster assignment but also for the number of clusters while allowing for varying cluster sizes. In this setup, we provide a method to create a long-term point trajectory graph with attractive and repulsive binary terms and outperform state-of-the-art methods based on spectral clustering on the FBMS-59 dataset and on the motion subtask of the VSB100 dataset.

****

Action Localization in Videos Through Context Walk

Khurram Soomro, Haroon Idrees, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3280-3288

This paper presents an efficient approach for localizing actions by learning contextual relations, in the form of relative locations between different video regions. We begin by over-segmenting the videos into supervoxels, which have the ability to preserve action boundaries and also reduce the complexity of the problem. Context relations are learned during training which capture displacements from all the supervoxels in a video to those belonging to foreground actions. Then, given a testing video, we select a supervoxel randomly and use the context information acquired during training to estimate the probability of each supervoxel belonging to the foreground action. The walk proceeds to a new supervoxel and the process is repeated for a few steps. This ``context walk'' generates a conditional distribution of an action over all the supervoxels. A Conditional Random Field is then used to find action proposals in the video, whose confidences are obtained using SVMs. We validated the proposed approach on several datasets and show that context in the form of relative displacements between supervoxels can be extremely useful for action localization. This also results in significantly fewer evaluations of the classifier, in sharp contrast to the alternate sliding window approaches.

****

RGB-W: When Vision Meets Wireless

Alexandre Alahi, Albert Haque, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3289-3297

Inspired by the recent success of RGB-D cameras, we propose the enrichment of RGB data with an additional "quasi-free" modality, namely, the wireless signal (e.g., wifi or Bluetooth) emitted by individuals' cell phones, referred to as RGB-W. The received signal strength acts as a rough proxy for depth and a reliable cue on their identity. Although the measured signals are highly noisy (more than 2m average localization error), we demonstrate that the combination of visual and wireless data significantly improves the localization accuracy.   We introduce a novel image-driven representation of wireless data which embeds all received signals onto a single image. We then indicate the ability of this additional data to (i) locate persons within a sparsity-driven framework and to (ii) track individuals with a new confidence measure on the data association problem. Our solut ion outperforms existing localization methods by a significant margin. It can be applied to the millions of currently installed RGB cameras to better analyze human behavior and offer the next generation of high-accuracy location-based services.

****

Action Detection by Implicit Intentional Motion Clustering

Wei Chen, Jason J. Corso; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3298-3306

Explicitly using human detection and pose estimation has found limited success in action recognition problems. This may be due to the complexity in the articulated motion human exhibit. Yet, we know that action requires an actor and intention. This paper hence seeks to understand the spatiotemporal properties of intentional movement and how to capture such intentional movement without relying on challenging human detection and tracking. We conduct a quantitative analysis of intentional movement, and our findings motivate a new approach for implicit intentional movement extraction that is based on spatiotemporal trajectory clustering by leveraging the properties of intentional movement. The intentional movement clusters are then used as action proposals for detection. Our results on three action detection benchmarks indicate the relevance of focusing on intentional movement for action detection; our method significantly outperforms the state of the art on the challenging MSR-II multi-action video benchmark.

****

Simultaneous Foreground Detection and Classification With Hybrid Features

Jaemyun Kim, Adin Ramirez Rivera, Byungyong Ryu, Oksam Chae; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3307-3315

In this paper, we propose a hybrid background model that relies on edge and non-edge features of the image to produce the model. We encode these features into a coding scheme, that we called Local Hybrid Pattern (LHP), that selectively models edges and non-edges features of each pixel. Furthermore, we model each pixel with an adaptive code dictionary to represent the background dynamism, and update it by adding stable codes and discarding unstable ones. We weight each code in the dictionary to enhance its description of the pixel it models. The foreground is detected as the incoming codes that deviate from the dictionary. We can detect (as foreground or background) and classify (as edge or inner region) each pixel simultaneously. We tested our proposed method in existing databases with promising results.

****

Training a Feedback Loop for Hand Pose Estimation

Markus Oberweger, Paul Wohlhart, Vincent Lepetit; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3316-3324

We propose an entirely data-driven approach to estimating the 3D pose of a hand given a depth image. We show that we can correct the mistakes made by a Convolutional Neural Network trained to predict an estimate of the 3D pose by using a feedback loop. The components of this feedback loop are also Deep Networks, optimized using training data. They remove the need for fitting a 3D model to the input data, which requires both a carefully designed fitting function and algorithm. We show that our approach outperforms state-of-the-art methods, and is efficient as our implementation runs at over 400 fps on a single GPU.

****

Opening the Black Box: Hierarchical Sampling Optimization for Estimating Human Hand Pose

Danhang Tang, Jonathan Taylor, Pushmeet Kohli, Cem Keskin, Tae-Kyun Kim, Jamie Shotton; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3325-3333

We address the problem of hand pose estimation, formulated as an inverse problem. Typical approaches optimize an energy function over pose parameters using a `black box' image generation procedure. This procedure knows little about either the relationships between the parameters or the form of the energy function. In this paper, we show that we can significantly improving upon black box optimization by exploiting high-level knowledge of the structure of the parameters and us ing a local surrogate energy function. Our new framework, hierarchical sampling optimization, consists of a sequence of predictors organized into a kinematic hierarchy. Each predictor is conditioned on its ancestors, and generates a set of samples over a subset of the pose parameters. The highly-efficient surrogate energy is used to select among samples. Having evaluated the full hierarchy, the partial pose samples are concatenated to generate a full-pose hypothesis. Several hypotheses are generated using the same procedure, and finally the original full energy function selects the best result. Experimental evaluation on three publically available datasets show that our method is particularly impressive in low-compute scenarios where it significantly outperforms all other state-of-the-art methods.

****

Panoptic Studio: A Massively Multiview System for Social Motion Capture

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, Yaser Sheikh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3334-3342

We present an approach to capture the 3D structure and motion of a group of people engaged in a social interaction. The core challenges in capturing social interactions are: (1) occlusion is functional and frequent; (2) subtle motion needs to be measured over a space large enough to host a social group; and (3) human appearance and configuration variation is immense. The Panoptic Studio is a system organized around the thesis that social interactions should be measured through the perceptual integration of a large variety of view points. We present a modularized system designed around this principle, consisting of integrated structural, hardware, and software innovations. The system takes, as input, 480 synchronized video streams of multiple people engaged in social activities, and produces, as output, the labeled time-varying 3D structure of anatomical landmarks on individuals in the space. The algorithmic contributions include a hierarchical approach for generating skeletal trajectory proposals, and an optimization framework for skeletal reconstruction with trajectory re-association.

****

Where to Buy It: Matching Street Clothing Photos in Online Shops

M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, Tamara L. Berg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3343-3351

In this paper, we define a new task, Exact Street to Shop, where our goal is to match a real-world example of a garment item to the same item in an online shop. This is an extremely challenging task due to visual differences between street photos (pictures of people wearing clothing in everyday uncontrolled settings) and online shop photos (pictures of clothing items on people, mannequins, or in isolation, captured by professionals in more controlled settings). We collect a new dataset for this application containing 404,683 shop photos collected from 25 different online retailers and 20,357 street photos, providing a total of 39,479 clothing item matches between street and shop photos. We develop three different methods for Exact Street to Shop retrieval, including two deep learning baseline methods, and a method to learn a similarity measure between the street and shop domains. Experiments demonstrate that our learned similarity significantly outperforms our baselines that use existing deep learning based representations.

****

Multi-Task Recurrent Neural Network for Immediacy Prediction

Xiao Chu, Wanli Ouyang, Wei Yang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3352-3360

In this paper, we propose to predict immediacy for interacting persons from still images. A complete immediacy set includes interactions, relative distance, body leaning direction and standing orientation. These measures are found to be related to the attitude, social relationship, social interaction, action, nationality, and religion of the communicators. A large-scale dataset with 10,000 images is constructed, in which all the immediacy measures and the human poses are annotated. We propose a rich set of immediacy representations that help to predict i mmediacy from imperfect 1-person and 2-person pose estimation results. A multi-task deep recurrent neural network is constructed to take the proposed rich immediacy representation as input and learn the complex relationship among immediacy predictions multiple steps of refinement. The effectiveness of the proposed approach is proved through extensive experiments on the large scale dataset.

****

Learning Complexity-Aware Cascades for Deep Pedestrian Detection

Zhaowei Cai, Mohammad Saberian, Nuno Vasconcelos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3361-3369

The design of complexity-aware cascaded detectors, combining features of very different complexities, is considered. A new cascade design procedure is introduced, by formulating cascade learning as the Lagrangian optimization of a risk that accounts for both accuracy and complexity. A boosting algorithm, denoted as complexity aware cascade training (CompACT), is then derived to solve this optimization. CompACT cascades are shown to seek an optimal trade-off between accuracy and complexity by pushing features of higher complexity to the later cascade stages, where only a few difficult candidate patches remain to be classified. This enables the use of features of vastly different complexities in a single detector. In result, the feature pool can be expanded to features previously impractical for cascade design, such as the responses of a deep convolutional neural network (CNN). This is demonstrated through the design of a pedestrian detector with a pool of features whose complexities span orders of magnitude. The resulting cascade generalizes the combination of a CNN with an object proposal mechanism: rather than a pre-processing stage, CompACT cascades seamlessly integrate CNNs in their stages. This enables state of the art performance on the Caltech and KITTI datasets, at fairly fast speeds.

****

Polarized 3D: High-Quality Depth Sensing With Polarization Cues

Achuta Kadambi, Vage Taamazyan, Boxin Shi, Ramesh Raskar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3370-3378

Coarse depth maps can be enhanced by using the shape information from polarization cues. We propose a framework to combine surface normals from polarization (hereafter polarization normals) with an aligned depth map. Polarization normals have not been used for depth enhancement before. This is because polarization normals suffer from physics-based artifacts, such as azimuthal ambiguity, refractive distortion and fronto-parallel signal degradation. We propose a framework to overcome these key challenges, allowing the benefits of polarization to be used to enhance depth maps. Our results demonstrate improvement with respect to state-of-the-art 3D reconstruction techniques.

****

Airborne Three-Dimensional Cloud Tomography

Aviad Levis, Yoav Y. Schechner, Amit Aides, Anthony B. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3379-3387

We seek to sense the three dimensional (3D) volumetric distribution of scatterers in a heterogenous medium. An important case study for such a medium is the atmosphere. Atmospheric contents and their role in Earth's radiation balance have significant uncertainties with regards to scattering components: aerosols and clouds. Clouds, made of water droplets, also lead to local effects as precipitation and shadows. Our sensing approach is computational tomography using passive multi-angular imagery. For light-matter interaction that accounts for multiple-scattering, we use the 3D radiative transfer equation as a forward model. Volumetric recovery by inverting this model suffers from a computational bottleneck on large scales, which include many unknowns. Steps taken make this tomography tractable, without approximating the scattering order or angle range.

****

Leave-One-Out Kernel Optimization for Shadow Detection

Tomas F. Yago Vicente, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3388-3396

The objective of this work is to detect shadows in images. We pose this as the problem of labeling image regions, where each region corresponds to a group of su perpixels. To predict the label of each region, we train a kernel Least-Squares SVM for separating shadow and non-shadow regions. The parameters of the kernel and the classifier are jointly learned to minimize the leave-one-out cross validation error. Optimizing the leave-one-out cross validation error is typically difficult, but it can be done efficiently in our framework. Experiments on two challenging shadow datasets, UCF and UIUC, show that our region classifier outperforms more complex methods. We further enhance the performance of the region classifier by embedding it in an MRF framework and adding pairwise contextual cues. This leads to a method that significantly outperforms the state-of-the-art.

****

Removing Rain From a Single Image via Discriminative Sparse Coding

Yu Luo, Yong Xu, Hui Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3397-3405

Visual distortions on images caused by bad weather conditions can have a negative impact on the performance of many outdoor vision systems. One often seen bad weather is rain which causes significant yet complex local intensity fluctuations in images. The paper aims at developing an effective algorithm to remove visual effects of rain from a single rainy image, i.e. separate the rain layer and the de-rained image layer from an rainy image. Built upon a non-linear generative model of rainy image, namely screen blend mode, we proposed a dictionary learning based algorithm for single image de-raining. The basic idea is to sparsely approximate the patches of two layers by very high discriminative codes over a learned dictionary with strong mutual exclusivity property. Such discriminative sparse codes lead to accurate separation of two layers from their non-linear composite. The experiments showed that the proposed method outperformed the existing single image de-raining methods on tested rain images.

****

Mutual-Structure for Joint Filtering

Xiaoyong Shen, Chao Zhou, Li Xu, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3406-3414

Previous joint/guided filters directly transfer the structural information in the reference image to the target one. In this paper, we first analyze its major drawback -- that is, there may be completely different edges in the two images. Simply passing all patterns to the target could introduce significant errors. To address this issue, we propose the concept of mutual-structure, which refers to the structural information that is contained in both images and thus can be safely enhanced by joint filtering, and an untraditional objective function that can be efficiently optimized to yield mutual structure. Our method results in necessary and important edge preserving, which greatly benefits depth completion, optical flow estimation, image enhancement, stereo matching, to name a few.

****

Photometric Stereo in a Scattering Medium

Zak Murez, Tali Treibitz, Ravi Ramamoorthi, David Kriegman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3415-3423

Photometric stereo is widely used for 3D reconstruction. However, its use in scattering media such as water, biological tissue and fog has been limited until now, because of forward scattered light from both the source and object, as well as light scattered back from the medium (backscatter). Here we make three contributions to address the key modes of light propagation, under the common single scattering assumption for dilute media. First, we show through extensive simulations that single-scattered light from a source can be approximated by a point light source with a single direction. This alleviates the need to handle light source blur explicitly. Next, we model the blur due to scattering of light from the object. We measure the object point-spread function and introduce a simple deconvolution method. Finally, we show how imaging fluorescence emission where available, eliminates the backscatter component and increases the signal-to-noise ratio. Experimental results in a water tank, with different concentrations of scattering media added, show that deconvolution produces higher-quality 3D reconstructions than previous techniques, and that when combined with fluorescence, can produce results similar to that in clear water even for highly turbid media.

****

Resolving Scale Ambiguity Via XSlit Aspect Ratio Analysis

Wei Yang, Haiting Lin, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3424-3432

In perspective cameras, images of a frontal-parallel 3D object preserve its aspect ratio invariant to its depth. Such an invariance is useful in photography but is unique to perspective projection. In this paper, we show that alternative non-perspective cameras such as the crossed-slit or XSlit cameras exhibit a different depth-dependent aspect ratio (DDAR) property that can be used to 3D recovery. We first conduct a comprehensive analysis to characterize DDAR, infer object depth from its AR, and model recoverable depth range, sensitivity, and error. We show that repeated shape patterns in real Manhattan World scenes can be used for 3D reconstruction using a single XSlit image. We also extend our analysis to model slopes of lines. Specifically, parallel 3D lines exhibit depth-dependent slopes (DDS) on their images which can also be used to infer their depths. We validate our analyses using real XSlit cameras, XSlit panoramas, and catadioptric mirrors. Experiments show that DDAR and DDS provide important depth cues and enable effective single-image scene reconstruction.

****

Single-Shot Specular Surface Reconstruction With Gonio-Plenoptic Imaging

Lingfei Meng, Liyang Lu, Noah Bedard, Kathrin Berkner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3433-3441

We present a gonio-plenoptic imaging system that realizes a single-shot shape measurement for specular surfaces. The system is comprised of a collimated illumination source and a plenoptic camera. Unlike a conventional plenoptic camera, our system captures the BRDF variation of the object surface in a single image in addition to the light field information from the scene, which allows us to recover very fine 3D structures of the surface. The shape of the surface is reconstructed based on the reflectance property of the material rather than the parallax between different views. Since only a single-shot is required to reconstruct the whole surface, our system is able to capture dynamic surface deformation in a video mode. We also describe a novel calibration technique that maps the light field viewing directions from the object space to subpixels on the sensor plane. The proposed system is evaluated using a concave mirror with known curvature, and is compared to a parabolic mirror scanning system as well as a multi-illumination photometric stereo approach based on simulations and experiments.

****

TransCut: Transparent Object Segmentation From a Light-Field Image

Yichao Xu, Hajime Nagahara, Atsushi Shimada, Rin-ichiro Taniguchi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3442-3450

The segmentation of transparent objects can be very useful in computer vision applications. However, because they borrow texture from their background and have a similar appearance to their surroundings, transparent objects are not handled well by regular image segmentation methods. We propose a method that overcomes these problems using the consistency and distortion properties of a light-field image. Graph-cut optimization is applied for the pixel labeling problem. The light-field linearity is used to estimate the likelihood of a pixel belonging to the transparent object or Lambertian background, and the occlusion detector is used to find the occlusion boundary. We acquire a light field dataset for the transparent object, and use this dataset to evaluate our method. The results demonstrate that the proposed method successfully segments transparent objects from the background.

****

Depth Recovery From Light Field Using Focal Stack Symmetry

Haiting Lin, Can Chen, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3451-3459

We describe a technique to recover depth from a light field (LF) using two proposed features of the LF focal stack. One feature is the property that non-occluding pixels exhibit symmetry along the focal depth dimension centered at the in-fo cus slice. The other is a data consistency measure based on analysis-by-synthesis, i.e., the difference between the synthesized focal stack given the hypothesized depth map and that from the LF. These terms are used in an iterative optimization framework to extract scene depth. Experimental results on real Lytro and Raytrix data demonstrate that our technique outperforms state-of-the-art solutions and is significantly more robust to noise and under-sampling.

****

Depth Map Estimation and Colorization of Anaglyph Images Using Local Color Prior and Reverse Intensity Distribution

W. Williem, Ramesh Raskar, In Kyu Park; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3460-3468

In this paper, we present a joint iterative anaglyph stereo matching and colorization framework for obtaining a set of disparity maps and colorized images. Conventional stereo matching algorithms fail when addressing anaglyph images that do not have similar intensities on their two respective view images. To resolve this problem, we propose two novel data costs using local color prior and reverse intensity distribution factor for obtaining accurate depth maps. To colorize an anaglyph image, each pixel in one view is warped to another view using the obtained disparity values of non-occluded regions. A colorization algorithm using optimization is then employed with additional constraint to colorize the remaining occluded regions. Experimental results confirm that the proposed unified framework is robust and produces accurate depth maps and colorized stereo images.

****

Learning Data-Driven Reflectance Priors for Intrinsic Image Decomposition

Tinghui Zhou, Philipp Krahenbuhl, Alexei A. Efros; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3469-3477

We propose a data-driven approach for intrinsic image decomposition, which is the process of inferring the confounding factors of reflectance and shading in an image. We pose this as a two-stage learning problem. First, we train a model to predict relative reflectance ordering be- tween image patches ('brighter', 'darker', 'same') from large-scale human annotations, producing a data-driven reflectance prior. Second, we show how to naturally integrate this learned prior into existing energy minimization frame- works for intrinsic image decomposition. We compare our method to the state-of-the-art approach of Bell et al. [7] on both decomposition and image relighting tasks, demonstrating the benefits of the simple relative reflectance prior, especially for scenes under challenging lighting conditions.

****

Photometric Stereo With Small Angular Variations

Jian Wang, Yasuyuki Matsushita, Boxin Shi, Aswin C. Sankaranarayanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3478-3486

Most existing successful photometric stereo setups require large angular variations in illumination directions, which results in acquisition rigs that have large spatial extent. For many applications, especially involving mobile devices, it is important that the device be spatially compact. This naturally implies smaller angular variations in the illumination directions. This paper studies the effect of small angular variations in illumination directions to photometric stereo. We explore both theoretical justification and practical issues in the design of a compact and portable photometric stereo device on which a camera is surrounded by a ring of point light sources. We first derive the relationship between the estimation error of surface normal and the baseline of the point light sources. Armed with this theoretical insight, we develop a small baseline photometric stereo prototype to experimentally examine the theory and its practicality.

****

Occlusion-Aware Depth Estimation Using Light-Field Cameras

Ting-Chun Wang, Alexei A. Efros, Ravi Ramamoorthi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3487-3495

Consumer-level and high-end light-field cameras are now widely available. Recent work has demonstrated practical methods for passive depth estimation from ligh t-field images. However, most previous approaches do not explicitly model occlusions, and therefore cannot capture sharp transitions around object boundaries. A common assumption is that a pixel exhibits photo-consistency when focused to its correct depth, i.e., all viewpoints converge to a single (Lambertian) point in the scene. This assumption does not hold in the presence of occlusions, making most current approaches unreliable precisely where accurate depth information is most important - at depth discontinuities.   In this paper, we develop a depth estimation algorithm that treats occlusion explicitly; the method also enables identification of occlusion edges, which may be useful in other applications. We show that, although pixels at occlusions do not preserve photo-consistency in general, they are still consistent in approximately half the viewpoints. Moreover, the line separating the two view regions (correct depth vs. occluder) has the same orientation as the occlusion edge has in the spatial domain. By treating these two regions separately, depth estimation can be improved. Occlusion predictions can also be computed and used for regularization. Experimental results show that our method outperforms current state-of-the-art light-field depth estimation algorithms, especially near occlusion boundaries.

****

Oriented Light-Field Windows for Scene Flow

Pratul P. Srinivasan, Michael W. Tao, Ren Ng, Ravi Ramamoorthi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3496-3504

2D spatial image windows are used for comparing pixel values in computer vision applications such as correspondence for optical flow and 3D reconstruction, bilateral filtering, and image segmentation. However, pixel window comparisons can suffer from varying defocus blur and perspective at different depths, and can also lead to a loss of precision. In this paper, we leverage the recent use of light-field cameras to propose alternative - oriented light-field windows that enable more robust and accurate pixel comparisons. For Lambertian surfaces focused to the correct depth, the 2D distribution of angular rays from a pixel remains consistent. We build on this idea to develop an oriented 4D light-field window that accounts for shearing (depth), translation (matching), and windowing. Our main application is to scene flow, a generalization of optical flow to the 3D vector field describing the motion of each point in the scene. We show significant benefits of oriented light-field windows over standard 2D spatial windows. We also demonstrate additional applications of oriented light-field windows for bilateral filtering and image segmentation.

****

Extended Depth of Field Catadioptric Imaging Using Focal Sweep

Ryunosuke Yokoya, Shree K. Nayar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3505-3513

Catadioptric imaging systems use curved mirrors to capture wide fields of view. However, due to the curvature of the mirror, these systems tend to have very limited depth of field (DOF), with the point spread function (PSF) varying dramatically over the field of view and as a function of scene depth. In recent years, focal sweep has been used extensively to extend the DOF of conventional imaging systems. It has been shown that focal sweep produces an integrated point spread function (IPSF) that is nearly space-invariant and depth-invariant, enabling the recovery of an extended depth of field (EDOF) image by deconvolving the captured focal sweep image with a single IPSF. In this paper, we use focal sweep to extend the DOF of a catadioptric imaging system. We show that while the IPSF is spatially varying when a curved mirror is used, it remains quasi depth-invariant over the wide field of view of the imaging system. We have developed a focal sweep system where mirrors of different shapes can be used to capture wide field of view EDOF images. In particular, we show experimental results using spherical and paraboloidal mirrors.

****

Intrinsic Depth: Improving Depth Transfer With Intrinsic Images

Naejin Kong, Michael J. Black; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3514-3522

We formulate the estimation of dense depth maps from video sequences as a proble m of intrinsic image estimation. Our approach synergistically integrates the estimation of multiple intrinsic images including depth, albedo, shading, optical flow, and surface contours. We build upon an example-based framework for depth estimation that uses label transfer from a database of RGB and depth pairs. We combine this with a method that extracts consistent albedo and shading from video. In contrast to raw RGB values, albedo and shading provide a richer, more physical, foundation for depth transfer. Additionally we train a new contour detector to predict surface boundaries from albedo, shading, and pixel values and use this to improve the estimation of depth boundaries. We also integrate sparse structure from motion with our method to improve the metric accuracy of the estimated depth maps. We evaluate our Intrinsic Depth method quantitatively by estimating depth from videos in the NYU RGB-D and SUN3D datasets. We find that combining the estimation of multiple intrinsic images improves depth estimation relative to the baseline method.

****

Separating Fluorescent and Reflective Components by Using a Single Hyperspectral Image

Yinqiang Zheng, Ying Fu, Antony Lam, Imari Sato, Yoichi Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3523-3531

This paper introduces a novel method to separate fluorescent and reflective components in the spectral domain. In contrast to existing methods, which require to capture two or more images under varying illuminations, we aim to achieve this separation task by using a single hyperspectral image. After identifying the critical hurdle in single-image component separation, we mathematically design the optimal illumination spectrum, which is shown to contain substantial high-frequency components in the frequency domain. This observation, in turn, leads us to recognize a key difference between reflectance and fluorescence in response to the frequency modulation effect of illumination, which fundamentally explains the feasibility of our method. On the practical side, we successfully find an off-the-shelf lamp as the light source, which is strong in irradiance intensity and cheap in cost. A fast linear separation algorithm is developed as well. Experiments using both synthetic data and real images have confirmed the validity of the selected illuminant and the accuracy of our separation algorithm.

****

Frequency-Based Environment Matting by Compressive Sensing

Yiming Qian, Minglun Gong, Yee-Hong Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3532-3540

Extracting environment mattes using existing approaches often requires either thousands of captured images or a long processing time, or both. In this paper, we propose a novel approach to capturing and extracting the matte of a real scene effectively and efficiently. Grown out of the traditional frequency-based signal analysis, our approach can accurately locate contributing sources. By exploiting the recently developed compressive sensing theory, we simplify the data acquisition process of frequency-based environment matting. Incorporating phase information in a frequency signal into data acquisition further accelerates the matte extraction procedure. Compared with the state-of-the-art method, our approach achieves superior performance on both synthetic and real data, while consuming only a fraction of the processing time.

****

Complementary Sets of Shutter Sequences for Motion Deblurring

Hae-Gon Jeon, Joon-Young Lee, Yudeog Han, Seon Joo Kim, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3541-3549

In this paper, we present a novel multi-image motion deblurring method utilizing the coded exposure technique. The key idea of our work is to capture video frames with a set of complementary fluttering patterns to preserve spatial frequency details. We introduce an algorithm for generating a complementary set of binary sequences based on the modern communication theory and implement the coded exposure video system with an off-the-shelf machine vision camera. The effectiveness of our method is demonstrated on various challenging examples with quantitative and qualitative comparisons to other computational image capturing methods used for image deblurring.

****

Hyperspectral Compressive Sensing Using Manifold-Structured Sparsity Prior

Lei Zhang, Wei Wei, Yanning Zhang, Fei Li, Chunhua Shen, Qinfeng Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3550-3558

To reconstruct hyperspectral image (HSI) accurately from a few noisy compressive measurements, we present a novel manifold-structured sparsity prior based hyperspectral compressive sensing (HCS) method in this study. A matrix based hierarchical prior is first proposed to represent the spectral structured sparsity and spatial unknown manifold structure of HSI simultaneously. Then, a latent variable Bayes model is introduced to learn the sparsity prior and estimate the noise jointly from measurements. The learned prior can fully represent the inherent 3D structure of HSI and regulate its shape based on the estimated noise level. Thus, with this learned prior, the proposed method improves the reconstruction accuracy significantly and shows strong robustness to unknown noise in HCS. Experiments on four real hyperspectral datasets show that the proposed method outperforms several state-of-the-art methods on the reconstruction accuracy of HSI.

****

A Gaussian Process Latent Variable Model for BRDF Inference

Stamatios Georgoulis, Vincent Vanweddingen, Marc Proesmans, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3559-3567

The problem of estimating a full BRDF from partial observations has already been studied using either parametric or non-parametric approaches. The goal in each case is to best match this sparse set of input measurements. In this paper we address the problem of inferring higher order reflectance information starting from the minimal input of a single BRDF slice. We begin from the prototypical case of a homogeneous sphere, lit by a head-on light source, which only holds information about less than 0.001% of the whole BRDF domain. We propose a novel method to infer the higher dimensional properties of the material's BRDF, based on the statistical distribution of known material characteristics observed in real-life samples. We evaluated our method based on a large set of experiments generated from real-world BRDFs and newly measured materials. Although inferring higher dimensional BRDFs from such modest training is not a trivial problem, our method performs better than state-of-the-art parametric, semi-parametric and non-parametric approaches. Finally, we discuss interesting applications on material re-lighting, and flash-based photography.

****

Active One-Shot Scan for Wide Depth Range Using a Light Field Projector Based on Coded Aperture

Hiroshi Kawasaki, Satoshi Ono, Yuki Horita, Yuki Shiba, Ryo Furukawa, Shinsaku Hiura; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3568-3576

The central projection model commonly used to model cameras as well as projectors, results in similar advantages and disadvantages in both types of system. Considering the case of active stereo systems using a projector and camera setup, a central projection model creates several problems; among them, narrow depth range and necessity of wide baseline are crucial. In the paper, we solve the problems by introducing a light field projector, which can project a depth-dependent pattern. The light field projector is realized by attaching a coded aperture with a high frequency mask in front of the lens of the video projector, which also projects a high frequency pattern. Because the light field projector cannot be approximated by a thin lens model and a precise calibration method is not established yet, an image-based approach is proposed to apply a stereo technique to the system. Although image-based techniques usually require a large database and often imply heavy computational costs, we propose a hierarchical approach and a feature-based search for solution. In the experiments, it is confirmed that our method can accurately recover the dense shape of curved and textured objects for a w ide range of depths from a single captured image.

****

Model-Based Tracking at 300Hz Using Raw Time-of-Flight Observations

Jan Stuhmer, Sebastian Nowozin, Andrew Fitzgibbon, Richard Szeliski, Travis Perry, Sunil Acharya, Daniel Cremers, Jamie Shotton; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3577-3585

Consumer depth cameras have dramatically improved our ability to track rigid, articulated, and deformable 3D objects in real-time. However, depth cameras have a limited temporal resolution (frame-rate) that restricts the accuracy and robustness of tracking, especially for fast or unpredictable motion. In this paper, we show how to perform model-based object tracking which allows to reconstruct the object's depth at an order of magnitude higher frame-rate through simple modifications to an off-the-shelf depth camera. We focus on phase-based time-of-flight (ToF) sensing, which reconstructs each low frame-rate depth image from a set of short exposure 'raw' infrared captures. These raw captures are taken in quick succession near the beginning of each depth frame, and differ in the modulation of their active illumination. We make two contributions. First, we detail how to perform model-based tracking against these raw captures. Second, we show that by reprogramming the camera to space the raw captures uniformly in time, we obtain a 10x higher frame-rate, and thereby improve the ability to track fast-moving objects.

****

Hyperspectral Super-Resolution by Coupled Spectral Unmixing

Charis Lanaras, Emmanuel Baltsavias, Konrad Schindler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3586-3594

Hyperspectral cameras capture images with many narrow spectral channels, which densely sample the electromagnetic spectrum. The detailed spectral resolution is useful for many image analysis problems, but it comes at the cost of much lower spatial resolution. Hyperspectral super-resolution addresses this problem, by fusing a low-resolution hyperspectral image and a conventional high-resolution image into a product of both high spatial and high spectral resolution. In this paper, we propose a method which performs hyperspectral super-resolution by jointly unmixing the two input images into the pure reflectance spectra of the observed materials and the associated mixing coefficients. The formulation leads to a coupled matrix factorisation problem, with a number of useful constraints imposed by elementary physical properties of spectral mixing. In experiments with two benchmark datasets we show that the proposed approach delivers improved hyperspectral super-resolution.

****

Depth Selective Camera: A Direct, On-Chip, Programmable Technique for Depth Selectivity in Photography

Ryuichi Tadano, Adithya Kumar Pediredla, Ashok Veeraraghavan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3595-3603

Time of flight (ToF) cameras use a temporally modulated light source and measure correlation between the reflected light and a sensor modulation pattern, in order to infer scene depth. In this paper, we show that such correlational sensors can also be used to selectively accept or reject light rays from certain scene depths. The basic idea is to carefully select illumination and sensor modulation patterns such that the correlation is non-zero only in the selected depth range - thus light reflected from objects outside this depth range do not affect the correlational measurements. We demonstrate a prototype depth-selective camera and highlight two potential applications: imaging through scattering media and virtual blue screening. This depthselectivity can be used to reject back-scattering and reflection from media in front of the subjects of interest, thereby significantly enhancing the ability to image through scattering media- critical for applications such as car navigation in fog and rain. Similarly, such depth selectivity can also be utilized as a virtual blue-screen in cinematography by rejecting light reflecting from background, while selectively retaining light contributions from the foreground subject.

****

A Groupwise Multilinear Correspondence Optimization for 3D Faces

Timo Bolkart, Stefanie Wuhrer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3604-3612

Multilinear face models are widely used to model the space of human faces with expressions. For databases of 3D human faces of different identities performing multiple expressions, these statistical shape models decouple identity and expression variations. To compute a high-quality multilinear face model, the quality of the registration of the database of 3D face scans used for training is essential. Meanwhile, a multilinear face model can be used as an effective prior to register 3D face scans, which are typically noisy and incomplete. Inspired by the minimum description length approach, we propose the first method to jointly optimize a multilinear model and the registration of the 3D scans used for training. Given an initial registration, our approach fully automatically improves the registration by optimizing an objective function that measures the compactness of the multilinear model, resulting in a sparse model. We choose a continuous representation for each face shape that allows to use a quasi-Newton method in parameter space for optimization. We show that our approach is computationally significantly more efficient and leads to correspondences of higher quality than existing methods based on linear statistical models. This allows us to evaluate our approach on large standard 3D face databases and in the presence of noisy initializations.

****

Selective Encoding for Recognizing Unreliably Localized Faces

Ang Li, Vlad Morariu, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3613-3621

Most existing face verification systems rely on precise face detection and registration. However, these two components are fallible under unconstrained scenarios (e.g., mobile face authentication) due to partial occlusions, pose variations, lighting conditions and limited view-angle coverage of mobile cameras. We address the unconstrained face verification problem by encoding face images directly without any explicit models of detection or registration. We propose a selective encoding framework which injects relevance information (e.g., foreground/background probabilities) into each cluster of a descriptor codebook. An additional selector component also discards distractive image patches and improves spatial robustness. We evaluate our framework using Gaussian mixture models and Fisher vectors on challenging face verification datasets. We apply selective encoding to Fisher vector features, which in our experiments degrade quickly with inaccurate face localization; our framework improves robustness with no extra test time computation. We also apply our approach to mobile based active face authentication task, demonstrating its utility in real scenarios.

****

Confidence Preserving Machine for Facial Action Unit Detection

Jiabei Zeng, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, Zhang Xiong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3622-3630

Varied sources of error contribute to the challenge of facial action unit detection. Previous approaches address specific and known sources. However, many sources are unknown. To address the ubiquity of error, we propose a Confident Preserving Machine (CPM) that follows an easy-to-hard classification strategy. During training, CPM learns two confident classifiers. A confident positive classifier separates easily identified positive samples from all else; a confident negative classifier does same for negative samples. During testing, CPM then learns a person-specific classifier using ``virtual labels'' provided by confident classifiers. This step is achieved using a quasi-semi-supervised (QSS) approach. Hard samples are typically close to the decision boundary, and the QSS approach disambiguates them using spatio-temporal constraints. To evaluate CPM, we compared it with a baseline single-margin classifier and state-of-the-art semi-supervised learning, transfer learning, and boosting methods in three datasets of spontaneous facial behavior. With few exceptions, CPM outperformed baseline and state-of-the art methods.

****

Learning Social Relation Traits From Face Images

Zhanpeng Zhang, Ping Luo, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3631-3639

Social relation defines the association, e.g., warm, friendliness, and dominance, between two or more people. Motivated by psychological studies, we investigate if such fine grained and high-level relation traits can be characterised and quantified from face images in the wild. To address this challenging problem we propose a deep model that learns a rich face representation to capture gender, expression, head pose, and age-related attributes, and then performs pairwise-face reasoning for relation prediction. To learn from heterogeneous attribute sources, we formulate a new network architecture with a bridging layer to leverage the inherent correspondences among these datasets. It can also cope with missing target attribute labels. Extensive experiments show that our approach is effective for fine-grained social relation learning in images and videos.

****

Robust Heart Rate Measurement From Video Using Select Random Patches

Antony Lam, Yoshinori Kuno; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3640-3648

The ability to remotely measure heart rate from videos without requiring any special setup is beneficial to many applications. In recent years, a number of papers on heart rate (HR) measurement from videos have been proposed. However, these methods typically require the human subject to be stationary and for the illumination to be controlled. For methods that do take into account motion and illumination changes, strong assumptions are still made about the environment (e.g. background can be used for illumination rectification). In this paper, we propose an HR measurement method that is robust to motion, illumination changes, and does not require use of an environment's background. We present conditions under which cardiac activity extraction from local regions of the face can be treated as a linear Blind Source Separation problem and propose a simple but robust algorithm for selecting good local regions. The independent HR estimates from multiple local regions are then combined in a majority voting scheme that robustly recovers the HR. We validate our algorithm on a large database of challenging videos.

****

Robust Model-Based 3D Head Pose Estimation

Gregory P. Meyer, Shalini Gupta, Iuri Frosio, Dikpal Reddy, Jan Kautz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3649-3657

We introduce a method for accurate three dimensional head pose estimation using a commodity depth camera. We perform pose estimation by registering a morphable face model to the measured depth data, using a combination of particle swarm optimization (PSO) and the iterative closest point (ICP) algorithm, which minimizes a cost function that includes a 3D registration and a 2D overlap term. The pose is estimated on the fly without requiring an explicit initialization or training phase. Our method handles large pose angles and partial occlusions by dynamically adapting to the reliable visible parts of the face. It is robust and generalizes to different depth sensors without modification. On the Biwi Kinect dataset, we achieve best-in-class performance, with average angular errors of 2.1, 2.1 and 2.4 degrees for yaw, pitch, and roll, respectively, and an average translational error of 5.9 mm, while running at 6 fps on a graphics processing unit.

****

Robust Facial Landmark Detection Under Significant Head Poses and Occlusion

Yue Wu, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3658-3666

There have been tremendous improvements for facial landmark detection on general "in-the-wild" images. However, it is still challenging to detect the facial landmarks on images with severe occlusion and images with large head poses (e.g. profile face). In fact, the existing algorithms usually can only handle one of them. In this work, we propose a unified robust cascade regression framework that can handle both images with severe occlusion and images with large head poses. Sp ecifically, the method iteratively predicts the landmark occlusions and the landmark locations. For occlusion estimation, instead of directly predicting the binary occlusion vectors, we introduce a supervised regression method that gradually updates the landmark visibility probabilities in each iteration to achieve robustness. In addition, we explicitly add occlusion pattern as a constraint to improve the performance of occlusion prediction. For landmark detection, we combine the landmark visibility probabilities, the local appearances, and the local shapes to iteratively update their positions. The experimental results show that the proposed method is significantly better than state-of-the-art works on images with severe occlusion and images with large head poses. It is also comparable to other methods on general "in-the-wild" images.

****

Conditional Convolutional Neural Network for Modality-Aware Face Recognition

Chao Xiong, Xiaowei Zhao, Danhang Tang, Karlekar Jayashree, Shuicheng Yan, Tae-Kyun Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3667-3675

Faces in the wild are usually captured with various poses, illuminations and occlusions, and thus inherently multimodally distributed in many tasks. We propose a conditional Convolutional Neural Network, named as c-CNN, to handle multimodal face recognition. Different from traditional CNN that adopts fixed convolution kernels, samples in c-CNN are processed with dynamically activated sets of kernels. In particular, convolution kernels within each layer are only sparsely activated when a sample is passed through the network. For a given sample, the activations of convolution kernels in a certain layer are conditioned on its present intermediate representation and the activation status in the lower layers. The activated kernels across layers define the sample-specific adaptive routes that reveal the distribution of underlying modalities. Consequently, the proposed framework does not rely on any prior knowledge of modalities in contrast with most existing methods. To substantiate the generic framework, we introduce a special case of c-CNN via incorporating the conditional routing of the decision tree, which is evaluated with two problems of multimodality - multi-view face identification and occluded face verification. Extensive experiments demonstrate consistent improvements over the counterparts unaware of modalities.

****

From Facial Parts Responses to Face Detection: A Deep Learning Approach

Shuo Yang, Ping Luo, Chen-Change Loy, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3676-3684

In this paper, we propose a novel deep convolutional network (DCN) that achieves outstanding performance on FDDB, PASCAL Face, and AFW. Specifically, our method achieves a high recall rate of 90.99% on the challenging FDDB benchmark, outperforming the state-of-the-art method by a large margin of 2.91%. Importantly, we consider finding faces from a new perspective through scoring facial parts responses by their spatial structure and arrangement. The scoring mechanism is carefully formulated considering challenging cases where faces are only partially visible. This consideration allows our network to detect faces under severe occlusion and unconstrained pose variation, which are the main difficulty and bottleneck of most existing face detection approaches. We show that despite the use of DCN, our network can achieve practical runtime speed.

****

Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification

Shengcai Liao, Stan Z. Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3685-3693

Person re-identification is becoming a hot research topic due to its value in both machine learning research and video surveillance applications. For this challenging problem, distance metric learning is shown to be effective in matching person images. However, existing approaches either require a heavy computation due to the positive semidefinite (PSD) constraint, or ignore the PSD constraint and learn a free distance function that makes the learned metric potentially noisy. We argue that the PSD constraint provides a useful regularization to smooth the solution of the metric, and hence the learned metric is more robust than without the PSD constraint. Another problem with metric learning algorithms is that the number of positive sample pairs is very limited, and the learning process is largely dominated by the large amount of negative sample pairs. To address the above issues, we derive a logistic metric learning approach with the PSD constraint and an asymmetric sample weighting strategy. Besides, we successfully apply the accelerated proximal gradient approach to find a global minimum solution of the proposed formulation, with a convergence rate of O(1/t^2) where t is the number of iterations. The proposed algorithm termed MLAPG is shown to be computationally efficient and able to perform low rank selection. We applied the proposed method for person re-identification, achieving state-of-the-art performance on four challenging databases (VIPeR, QMUL GRID, CUHK Campus, and CUHK03), compared to existing metric learning methods as well as published results.

****

Pose-Invariant 3D Face Alignment

Amin Jourabloo, Xiaoming Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3694-3702

Face alignment aims to estimate the locations of a set of landmarks for a given image.This problem has received much attention as evidenced by the recent advancement in both the methodology and performance. However, most of the existing works neither explicitly handle face images with arbitrary poses, nor perform large-scale experiments on non-frontal and profile face images. In order to address these limitations, this paper proposes a novel face alignment algorithm that estimates both 2D and 3D landmarks and their 2D visibilities for a face image with an arbitrary pose. By integrating a 3D point distribution model, a cascaded coupled-regressor approach is designed to estimate both the camera projection matrix and the 3D landmarks. Furthermore, the 3D model also allows us to automatically estimate the 2D landmark visibilities via surface normal. We use a substantially larger collection of all-pose face images to evaluate our algorithm and demonstrate superior performances than the state-of-the-art methods.

****

From Emotions to Action Units With Hidden and Semi-Hidden-Task Learning

Adria Ruiz, Joost Van de Weijer, Xavier Binefa; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3703-3711

Limited annotated training data is a challenging problem in Action Unit recognition. In this paper, we investigate how the use of large databases labelled according to the 6 universal facial expressions can increase the generalization ability of Action Unit classifiers. For this purpose, we propose a novel learning framework: Hidden-Task Learning. HTL aims to learn a set of Hidden-Tasks (Action Units) for which samples are not available but, in contrast, training data is easier to obtain from a set of related Visible-Tasks (Facial Expressions). To that end, HTL is able to exploit prior knowledge about the relation between Hidden and Visible-Tasks. In our case, we base this prior knowledge on empirical psychological studies providing statistical correlations between Action Units and universal facial expressions. Additionally, we extend HTL to Semi-Hidden Task Learning (SHTL) assuming that Action Unit training samples are also provided. Performing exhaustive experiments over four different datasets, we show that HTL and SHTL improve the generalization ability of AU classifiers by training them with additional facial expression data. Additionally, we show that SHTL achieves competitive performance compared with state-of-the-art Transductive Learning approaches which face the problem of limited training data by using unlabelled test samples during training.

****

Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face

Jungseock Joo, Francis F. Steen, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3712-3720

The human face is a primary medium of human communication and a prominent source of information used to infer various attributes. In this paper, we study a fully automated system that can infer the perceived traits of a person from his face

-- social dimensions, such as "intelligence," "honesty," and "competence" -- and how those traits can be used to predict the outcomes of real-world social events that involve long-term commitments, such as political elections, job hires, and marriage engagements. To this end, we propose a hierarchical model for enduring traits inferred from faces, incorporating high-level perceptions and intermediate-level attributes. We show that our trained model can successfully classify the outcomes of two important political events, only using the photographs of politicians' faces. Firstly, it classifies the winners of a series of recent U.S. elections with the accuracy of 67.9% (Governors) and 65.5% (Senators). We also reveal that the different political offices require different types of preferred traits. Secondly, our model can categorize the political party affiliations of politicians, i.e., Democrats vs. Republicans, with the accuracy of 62.6% (male) and 60.1% (female). To the best of our knowledge, our paper is the first to use automated visual trait analysis to predict the outcomes of real-world social events. This approach is more scalable and objective than the prior behavioral studies, and opens for a range of new applications.
****
Simultaneous Local Binary Feature Learning and Encoding for Face Recognition
Jiwen Lu, Venice Erin Liong, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3721-3729
In this paper, we propose a simultaneous local binary feature learning and encoding (SLBFLE) method for face recognition. Different from existing hand-crafted face descriptors such as local binary pattern (LBP) and Gabor features which require strong prior knowledge, our SLBFLE is an unsupervised feature learning approach which is automatically learned from raw pixels. Unlike existing binary face descriptors such as the LBP and discriminant face descriptor (DFD) which use a two-stage feature extraction approach, our SLBFLE jointly learns binary codes for local face patches and the codebook for feature encoding so that discriminative information from raw pixels can be simultaneously learned with a one-stage procedure. Experimental results on four widely used face datasets including LFW, YouTube Face (YTF), FERET and PaSC clearly demonstrate the effectiveness of the proposed method.
****
Deep Learning Face Attributes in the Wild
Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3730-3738
Predicting face attributes in the wild is challenging due to complex face variations. We propose a novel deep learning framework for attribute prediction in the wild. It cascades two CNNs, LNet and ANet, which are fine-tuned jointly with attribute tags, but pre-trained differently. LNet is pre-trained by massive general object categories for face localization, while ANet is pre-trained by massive face identities for attribute prediction. This framework not only outperforms the state-of-the-art with a large margin, but also reveals valuable facts on learning face representation. (1) It shows how the performances of face localization (LNet) and attribute prediction (ANet) can be improved by different pre-training strategies. (2) It reveals that although the filters of LNet are fine-tuned only with image-level attribute tags, their response maps over entire images have strong indication of face locations. This fact enables training LNet for face localization with only image-level annotations, but without face bounding boxes or landmarks, which are required by all attribute recognition works. (3) It also demonstrates that the high-level hidden neurons of ANet automatically discover semantic concepts after pre-training with massive face identities, and such concepts are significantly enriched after fine-tuning with attribute tags. Each attribute can be well explained with a sparse linear combination of these concepts.
****
Multi-Task Learning With Low Rank Attribute Embedding for Person Re-Identification
Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S. Davis, Wen Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3739-3747

We propose a novel Multi-Task Learning with Low Rank Attribute Embedding (MTL-LORAE) framework for person re-identification. Re-identifications from multiple cameras are regarded as related tasks to exploit shared information to improve re-identification accuracy. Both low level features and semantic/data-driven attributes are utilized. Since attributes are generally correlated, we introduce a low rank attribute embedding into the MTL formulation to embed original binary attributes to a continuous attribute space, where incorrect and incomplete attributes are rectified and recovered to better describe people. The learning objective function consists of a quadratic loss regarding class labels and an attribute embedding error, which is solved by an alternating optimization procedure. Experiments on three person re-identification datasets have demonstrated that MTL-LORAE outperforms existing approaches by a large margin and produces state-of-the-art results.

****

Regressing a 3D Face Shape From a Single Image

Sergey Tulyakov, Nicu Sebe; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3748-3755

In this work we present a method to estimate a 3D face shape from a single image. Our method is based on a cascade regression framework that directly estimates face landmarks locations in 3D. We include the knowledge that a face is a 3D object into the learning pipeline and show how this information decreases localization errors while keeping the computational time low. We predict the actual positions of the landmarks even if they are occluded due to face rotation. To support the ability of our method to reliably reconstruct 3D shapes, we introduce a simple method for head pose estimation using a single image that reaches higher accuracy than the state of the art. Comparison of 3D face landmarks localization with the available state of the art further supports the feasibility of a single-step face shape estimation. The code, trained models and our 3D annotations will be made available to the research community.

****

Rendering of Eyes for Eye-Shape Registration and Gaze Estimation

Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, Andreas Bulling; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3756-3764

Images of the eye are key in several computer vision problems, such as shape registration and gaze estimation. Recent large-scale supervised methods for these problems require time-consuming data collection and manual annotation, which can be unreliable. We propose synthesizing perfectly labelled photo-realistic training data in a fraction of the time. We used computer graphics techniques to build a collection of dynamic eye-region models from head scan geometry. These were randomly posed to synthesize close-up eye images for a wide range of head poses, gaze directions, and illumination conditions. We used our model's controllability to verify the importance of realistic illumination and shape variations in eye-region training data. Finally, we demonstrate the benefits of our synthesized training data (SynthesEyes) by out-performing state-of-the-art methods for eye-shape registration as well as cross-dataset appearance-based gaze estimation in the wild.

****

Multi-Scale Learning for Low-Resolution Person Re-Identification

Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, Shaogang Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3765-3773

In real world person re-identification (re-id), images of people captured at very different resolutions from different locations need be matched. Existing re-id models typically normalise all person images to the same size. However, a low-resolution (LR) image contains much less information about a person, and direct image scaling and simple size normalisation as done in conventional re-id methods cannot compensate for the loss of information. To solve this LR person re-id problem, we propose a novel joint multi-scale learning framework, termed joint multi-scale discriminant component analysis (JUDEA). The key component of this fram ework is a heterogeneous class mean discrepancy (HCMD) criterion for cross-scale image domain alignment, which is optimised simultaneously with discriminant modelling across multiple scales in the joint learning framework. Our experiments show that the proposed JUDEA framework outperforms existing representative re-id methods as well as other related LR visual matching models applied for the LR person re-id problem.

****

Learning to Transfer: Transferring Latent Task Structures and Its Application to Person-Specific Facial Action Unit Detection

Timur Almaev, Brais Martinez, Michel Valstar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3774-3782

In this article we explore the problem of constructing person-specific models for the detection of facial Action Units (AUs), addressing the problem from the point of view of Transfer Learning and Multi-Task Learning. Our starting point is the fact that some expressions, such as smiles, are very easily elicited, annotated, and automatically detected, while others are much harder to elicit and to annotate. We thus consider a novel problem: all AU models for the target subject are to be learnt using person-specific annotated data for a reference AU (AU12 in our case), and no data or little data regarding the target AU. In order to design such a model, we propose a novel Multi-Task Learning and the associated Transfer Learning framework, in which we consider both relations across subjects and AUs. That is to say, we consider a tensor structure among the tasks. Our approach hinges on learning the latent relations among tasks using one single reference AU, and then transferring these latent relations to other AUs. We show that we are able to effectively make use of the annotated data for AU12 when learning other person-specific AU models, even in the absence of data for the target task. Finally, we show the excellent performance of our method when small amounts of annotated data for the target tasks are made available.

****

Pairwise Conditional Random Forests for Facial Expression Recognition

Arnaud Dapogny, Kevin Bailly, Severine Dubuisson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3783-3791

Facial expression can be seen as the dynamic variation of one's appearance over time. Successful recognition thus involves finding representations of high-dimensional spatiotemporal patterns that can be generalized to unseen facial morphologies and variations of the expression dynamics. In this paper, we propose to learn Random Forests from heterogeneous derivative features (e.g. facial fiducial point movements or texture variations) upon pairs of images. Those forests are conditioned on the expression label of the first frame to reduce the variability of the ongoing expression transitions. When testing on a specific frame of a video, pairs are created between this frame and the previous ones. Predictions for each previous frame are used to draw trees from Pairwise Conditional Random Forests (PCRF) whose pairwise outputs are averaged over time to produce robust estimates. As such, PCRF appears as a natural extension of Random Forests to learn spatio-temporal patterns, that leads to significant improvements over standard Random Forests as well as state-of-the-art approaches on several facial expression benchmarks.

****

Multi-Conditional Latent Variable Model for Joint Facial Action Unit Detection

Stefanos Eleftheriadis, Ognjen Rudovic, Maja Pantic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3792-3800

We propose a novel multi-conditional latent variable model for simultaneous facial feature fusion and detection of facial action units. In our approach we exploit the structure-discovery capabilities of generative models such as Gaussian processes, and the discriminative power of classifiers such as logistic function. This leads to superior performance compared to existing classifiers for the target task that exploit either the discriminative or generative property, but not both. The model learning is performed via an efficient, newly proposed Bayesian learning strategy based on Monte Carlo sampling. Consequently, the learned model is robust to data overfitting, regardless of the number of both input features a nd jointly estimated facial action units. Extensive qualitative and quantitative experimental evaluations are performed on three publicly available datasets (CK+, Shoulder-pain and DISFA). We show that the proposed model outperforms the state-of-the-art methods for the target task on (i) feature fusion, and (ii) multiple facial action unit detection.

****

Leveraging Datasets With Varying Annotations for Face Alignment via Deep Regression Network

Jie Zhang, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3801-3809

Facial landmark detection, as a vital topic in computer vision, has been studied for many decades and lots of datasets have been collected for evaluation. These datasets usually have different annotations, e.g., 68-landmark markup for LFPW dataset, while 74-landmark markup for GTAV dataset. Intuitively, it is meaningful to fuse all the datasets to predict a union of all types of landmarks from multiple datasets (i.e., transfer the annotations of each dataset to all other datasets), but this problem is nontrivial due to the distribution discrepancy between datasets and incomplete annotations of all types for each dataset. In this work, we propose a deep regression network coupled with sparse shape regression (DRN-SSR) to predict the union of all types of landmarks by leveraging datasets with varying annotations, each dataset with one type of annotation. Specifically, the deep regression network intends to predict the union of all landmarks, and the sparse shape regression attempts to approximate those undefined landmarks on each dataset so as to guide the learning of the deep regression network for face alignment. Extensive experiments on two challenging datasets, IBUG and GLF, demonstrate that our method can effectively leverage the multiple datasets with different annotations to predict the union of all types of landmarks.

****

A Spatio-Temporal Appearance Representation for Viceo-Based Pedestrian Re-Identification

Kan Liu, Bingpeng Ma, Wei Zhang, Rui Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3810-3818

Pedestrian re-identification is a difficult problem due to the large variations in a person's appearance caused by different poses and viewpoints, illumination changes, and occlusions. Spatial alignment is commonly used to address these issues by treating the appearance of different body parts independently. However, a body part can also appear differently during different phases of an action. In this paper we consider the temporal alignment problem, in addition to the spatial one, and propose a new approach that takes the video of a walking person as input and builds a spatio-temporal appearance representation for pedestrian re-identification. Particularly, given a video sequence we exploit the periodicity exhibited by a walking person to generate a spatio-temporal body-action model, which consists of a series of body-action units corresponding to certain action primitives of certain body parts. Fisher vectors are learned and extracted from individual body-action units and concatenated into the final representation of the walking person. Unlike previous spatio-temporal features that only take into account local dynamic appearance information, our representation aligns the spatio-temporal appearance of a pedestrian globally. Extensive experiments on public datasets show the effectiveness of our approach compared with the state of the art.

****

Two Birds, One Stone: Jointly Learning Binary Code for Large-Scale Face Image Retrieval and Attributes Prediction

Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3819-3827

We address the challenging large-scale content-based face image retrieval problem, intended as searching images based on the presence of specific subject, given one face image of him/her. To this end, one natural demand is a supervised binary code learning method. While the learned codes might be discriminating, people often have a further expectation that whether some semantic message (e.g., visu al attributes) can be read from the human-incomprehensible codes. For this purpose, we propose a novel binary code learning framework by jointly encoding identity discriminability and a number of facial attributes into unified binary code. In this way, the learned binary codes can be applied to not only fine-grained face image retrieval, but also facial attributes prediction, which is the very innovation of this work, just like killing two birds with one stone. To evaluate the effectiveness of the proposed method, extensive experiments are conducted on a new purified large-scale web celebrity database, named CFW 60K, with abundant manual identity and attributes annotation, and experimental results exhibit the superiority of our method over state-of-the-art.

****

An Accurate Iris Segmentation Framework Under Relaxed Imaging Constraints Using Total Variation Model

Zijing Zhao, Kumar Ajay; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3828-3836

This paper proposes a novel and more accurate iris segmentation framework to automatically segment iris region from the face images acquired with relaxed imaging under visible or near-infrared illumination, which provides strong feasibility for applications in surveillance, forensics and the search for missing children, etc. The proposed framework is built on a novel total-variation based formulation which uses l1 norm regularization to robustly suppress noisy texture pixels for the accurate iris localization. A series of novel and robust post processing operations are introduced to more accurately localize the limbic boundaries. Our experimental results on three publicly available databases, i.e., FRGC, UBIRIS.v2 and CASIA.v4-distance, achieve significant performance improvement in terms of iris segmentation accuracy over the state-of-the-art approaches in the literature. Besides, we have shown that using iris masks generated from the proposed approach helps to improve iris recognition performance as well. Unlike prior work, all the implementations in this paper are made publicly available to further advance research and applications in biometrics at-d-distance.

****

Discriminative Pose-Free Descriptors for Face and Object Matching

Soubhik Sanyal, Sivaram Prasad Mudunuri, Soma Biswas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3837-3845

Pose invariant matching is a very important and challenging problem with various applications like recognizing faces in uncontrolled scenarios, matching objects taken from different view points, etc. In this paper, we propose a discriminative pose-free descriptor (DPFD) which can be used to match faces/objects across pose variations. Training examples at very few representative poses are used to generate virtual intermediate pose subspaces. An image or image region is then represented by a feature set obtained by projecting it on all these subspaces and a discriminative transform is applied on this feature set to make it suitable for classification tasks. Finally, this discriminative feature set is represented by a single feature vector, termed as DPFD. The DPFD of images taken from different viewpoints can be directly compared for matching. Extensive experiments on recognizing faces across pose, pose and resolution on the Multi-PIE and Surveillance Cameras Face datasets and comparisons with state-of-the-art approaches show the effectiveness of the proposed approach. Experiments on matching general objects across viewpoints show the generalizability of the proposed approach beyond faces.

****

Bi-Shifting Auto-Encoder for Unsupervised Domain Adaptation

Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3846-3854

In many real-world applications, the domain of model learning (referred as source domain) is usually inconsistent with or even different from the domain of testing (referred as target domain), which makes the learnt model degenerate in target domain, i.e., the test domain. To alleviate the discrepancy between source and target domains, we propose a domain adaptation method, named as Bi-shifting Auto-Encoder network (BAE). The proposed BAE attempts to shift source domain sampl es to target domain, and also shift the target domain samples to source domain. The non-linear transformation of BAE ensures the feasibility of shifting between domains, and the distribution consistency between the shifted domain and the desirable domain is constrained by sparse reconstruction between them. As a result, the shifted source domain is supervised and follows similar distribution as target domain. Therefore, any supervised method can be applied on the shifted source domain to train a classifier for classification in target domain. The proposed method is evaluated on three domain adaptation scenarios of face recognition, i.e., domain adaptation across view angle, ethnicity, and imaging sensor, and the promising results demonstrate that our proposed BAE can shift samples between domains and thus effectively deal with the domain discrepancy.

****

Regressive Tree Structured Model for Facial Landmark Localization

Gee-Sern Hsu, Kai-Hsiang Chang, Shih-Chieh Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3855-3861

Although the Tree Structured Model (TSM) is proven effective for solving face detection, pose estimation and landmark localization in an unified model, its sluggish run time makes it unfavorable in practical applications, especially when dealing with cases of multiple faces. We propose the Regressive Tree Structure Model (RTSM) to improve the run-time speed and localization accuracy. The RTSM is composed of two component TSMs, the coarse TSM (c-TSM) and the refined TSM (r-TSM), and a Bilateral Support Vector Regressor (BSVR). The c-TSM is built on the low-resolution octaves of samples so that it provides coarse but fast face detection. The r-TSM is built on the mid-resolution octaves so that it can locate the landmarks on the face candidates given by the c-TSM and improve precision. The r-TSM based landmarks are used in the forward BSVR as references to locate the dense set of landmarks, which are then used in the backward BSVR to relocate the landmarks with large localization errors. The forward and backward regression goes on iteratively until convergence. The performance of the RTSM is validated on three benchmark databases, the Multi-PIE, LFPW and AFW, and compared with the latest TSM to demonstrate its efficacy.

****

Person Recognition in Personal Photo Collections

Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3862-3870

Recognising persons in everyday photos presents major challenges (occluded faces, different clothing, locations, etc.) for machine vision. We propose a convnet based person recognition system on which we provide an in-depth analysis of informativeness of different body cues, impact of training data, and the common failure modes of the system. In addition, we discuss the limitations of existing benchmarks and propose more challenging ones. Our method is simple and is built on open source and open data, yet it improves the state of the art results on a large dataset of social media photos (PIPA).

****

Robust Statistical Face Frontalization

Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, Maja Pantic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3871-3879

Recently, it has been shown that excellent results can be achieved in both facial landmark localization and pose-invariant face recognition. These breakthroughs are attributed to the efforts of the community to manually annotate facial images in many different poses and to collect 3D facial data. In this paper, we propose a novel method for joint frontal view reconstruction and landmark localization using a small set of frontal images only. By observing that the frontal facial image is the one having the minimum rank of all different poses, an appropriate model which is able to jointly recover the frontalized version of the face as well as the facial landmarks is devised. To this end, a suitable optimization problem, involving the minimization of the nuclear norm and the matrix l1 norm is solved. The proposed method is assessed in frontal face reconstruction, face landmark localization, pose-invariant face recognition, and face verification in un constrained conditions. The relevant experiments have been conducted on 8 databases. The experimental results demonstrate the effectiveness of the proposed method in comparison to the state-of-the-art methods for the target problems.

****

PIEFA: Personalized Incremental and Ensemble Face Alignment

Xi Peng, Shaoting Zhang, Yu Yang, Dimitris N. Metaxas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3880-3888

Face alignment, especially on real-time or large-scale sequential images, is a challenging task with broad applications. Both generic and joint alignment approaches have been proposed with varying degrees of success. However, many generic methods are heavily sensitive to initializations and usually rely on offline-trained static models, which limit their performance on sequential images with extensive variations. On the other hand, joint methods are restricted to offline applications, since they require all frames to conduct batch alignment. To address these limitations, we propose to exploit incremental learning for personalized ensemble alignment. We sample multiple initial shapes to achieve image congealing within one frame, which enables us to incrementally conduct ensemble alignment by group-sparse regularized rank minimization. At the same time, personalized modeling is obtained by subspace adaptation under the same incremental framework, while correction strategy is used to alleviate model drifting. Experimental results on multiple controlled and in-the-wild databases demonstrate the superior performance of our approach compared with state-of-the-arts in terms of fitting accuracy and efficiency.

****

Understanding Everyday Hands in Action From RGB-D Images

Gregory Rogez, James S. Supancic III, Deva Ramanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3889-3897

We analyze functional manipulations of handheld objects, formalizing the problem as one of fine-grained grasp classification. To do so, we make use of a recently developed fine-grained taxonomy of human-object grasps. We introduce a large dataset of 12000 RGB-D images covering 71 everyday grasps in natural interactions. Our dataset is different from past work (typically addressed from a robotics perspective) in terms of its scale, diversity, and combination of RGB and depth data. From a computer-vision perspective, our dataset allows for exploration of contact and force prediction (crucial concepts in functional grasp analysis) from perceptual cues. We present extensive experimental results with state-of-the-art baselines, illustrating the role of segmentation, object context, and 3D-understanding in functional grasp analysis. We demonstrate a near 2X improvement over prior work and a naive deep baseline, while pointing out important directions for improvement.

****

Example-Based Modeling of Facial Texture From Deficient Data

Arnaud Dessein, William A. P. Smith, Richard C. Wilson, Edwin R. Hancock; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3898-3906

We present an approach to modeling ear-to-ear, high-quality texture from one or more partial views of a face with possibly poor resolution and noise. Our approach is example-based in that we reconstruct texture with patches from a database composed of previously seen faces. A 3D morphable model is used to establish shape correspondence between the observed data across views and training faces. The database is built on the mesh surface by segmenting it into uniform overlapping patches. Texture patches are selected by belief propagation so as to be consistent with neighbors and with observations in an appropriate image formation model. We also develop a variant that is insensitive to light and camera parameters, and incorporate soft symmetry constraints. We obtain textures of higher quality for degraded views as small as 10 pixels wide, than a standard model fitted to non-degraded data. We further show applications to super-resolution where we substantially improve quality compared to a state-of-the-art algorithm, and to texture completion where we fill in missing regions and remove facial clutter in a photorealistic manner.

****

Learning to Predict Saliency on Face Images

Mai Xu, Yun Ren, Zulin Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3907-3915

This paper proposes a novel method, which learns to detect saliency of face images. To be more specific, we obtain a database of eye tracking over extensive face images, via conducting an eye tracking experiment. With analysis on eye tracking database, we verify that the fixations tend to cluster around facial features, when viewing images with large faces. For modeling attention on faces and facial features, the proposed method learns the Gaussian mixture model (GMM) distribution from the fixations of eye tracking data as the top-down features for saliency detection of face images. Then, in our method, the top-down features (i.e., face and facial features) upon the the learnt GMM are linearly combined with the conventional bottom-up features (i.e., color, intensity, and orientation), for saliency detection. In the linear combination, we argue that the weights corresponding to top-down feature channels depend on the face size in images, and the relationship between the weights and face size is thus investigated via learning from the training eye tracking data. Finally, experimental results show that our learning-based method is able to advance state-of-the-art saliency prediction for face images. The corresponding database and code are available online: www.ee.buaa.edu.cn/xumfiles/saliency_detection.html.

****

Group Membership Prediction

Ziming Zhang, Yuting Chen, Venkatesh Saligrama; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3916-3924

The group membership prediction (GMP) problem involves predicting whether or not a collection of instances share a certain semantic property. For instance, in kinship verification given a collection of images, the goal is to predict whether or not they share a familial relationship. In this context we propose a novel probability model and introduce latent view-specific and view-shared random variables to jointly account for the view-specific appearance and cross-view similarities among data instances. Our model posits that data from each view is independent conditioned on the shared variables. This postulate leads to a parametric probability model that decomposes group membership likelihood into a tensor product of data-independent parameters and data-dependent factors. We propose learning the data-independent parameters in a discriminative way with bilinear classifiers, and test our prediction algorithm on challenging visual recognition tasks such as multi-camera person re-identification and kinship verification. On most benchmark datasets, our method can significantly outperform the current state-of-the-art.

****

Extraction of Virtual Baselines From Distorted Document Images Using Curvilinear Projection

Gaofeng Meng, Zuming Huang, Yonghong Song, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3925-3933

The baselines of a document page are a set of virtual horizontal and parallel lines, to which the printed contents of document, e.g., text lines, tables or inserted photos, are aligned. Accurate baseline extraction is of great importance in the geometric correction of curved document images. In this paper, we propose an efficient method for accurate extraction of these virtual visual cues from a curved document image. Our method comes from two basic observations that the baselines of documents do not intersect with each other and that within a narrow strip, the baselines can be well approximated by linear segments. Based upon these observations, we propose a curvilinear projection based method and model the estimation of curved baselines as a constrained sequential optimization problem. A dynamic programming algorithm is then developed to efficiently solve the problem. The proposed method can extract the complete baselines through each pixel of document images in a high accuracy. It is also scripts insensitive and highly robust to image noises, non-textual objects, image resolutions and image quality de gradation like blurring and non-uniform illumination. Extensive experiments on a number of captured document images demonstrate the effectiveness of the proposed method.

****

Robust RGB-D Odometry Using Point and Line Features

Yan Lu, Dezhen Song; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3934-3942

Lighting variation and uneven feature distribution are main challenges for indoor RGB-D visual odometry where color information is often combined with depth information. To meet the challenges, we fuse point and line features to form a robust odometry algorithm. Line features are abundant indoors and less sensitive to lighting change than points. We extract 3D points and lines from RGB-D data, analyze their measurement uncertainties, and compute camera motion using maximum likelihood estimation. We prove that fusing points and lines produces smaller motion estimate uncertainty than using either feature type alone. In experiments we compare our method with state-of-the-art methods including a keypoint-based approach and a dense visual odometry algorithm. Our method outperforms the counterparts under both constant and varying lighting conditions. Specifically, our method achieves an average translational error that is 34.9% smaller than the counterparts, when tested using public datasets.

****

Learning a Discriminative Model for the Perception of Realism in Composite Images

Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, Alexei A. Efros; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3943-3951

What makes an image appear realistic? In this work, we are answering this question from a data-driven perspective by learning the perception of visual realism directly from large amounts of data. In particular, we train a Convolutional Neural Network (CNN) model that distinguishes natural photographs from automatically generated composite images. The model learns to predict visual realism of a scene in terms of color, lighting and texture compatibility, without any human annotations pertaining to it. Our model outperforms previous works that rely on hand-crafted heuristics, for the task of classifying realistic vs. unrealistic photos. Furthermore, we apply our learned model to compute optimal parameters of a compositing method, to maximize the visual realism score predicted by our CNN model. We demonstrate its advantage against existing methods via a human perception study.

****

What Makes Tom Hanks Look Like Tom Hanks

Supasorn Suwajanakorn, Steven M. Seitz, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3952-3960

We reconstruct a controllable model of a person from a large photo collection that captures his or her persona, i.e., physical appearance and behavior. The ability to operate on unstructured photo collections enables modeling a huge number of people, including celebrities and other well photographed people without requiring them to be scanned. Moreover, we show the ability to drive or puppeteer the captured person B using any other video of a different person A. In this scenario, B acts out the role of person A, but retains his/her own personality and character. Our system is based on a novel combination of 3D face reconstruction, tracking, alignment, and multi-texture modeling, applied to the puppeteering problem. We demonstrate convincing results on a large variety of celebrities derived from Internet imagery and video.

****

Wide-Area Image Geolocalization With Aerial Reference Imagery

Scott Workman, Richard Souvenir, Nathan Jacobs; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3961-3969

We propose to use deep convolutional neural networks to address the problem of cross-view image geolocalization, in which the geolocation of a ground-level query image is estimated by matching to georeferenced aerial images. We use state-of

-the-art feature representations for ground-level images and introduce a cross-view training approach for learning a joint semantic feature representation for aerial images. We also propose a network architecture that fuses features extracted from aerial images at multiple spatial scales. To support training these networks, we introduce a massive database that contains pairs of aerial and ground-level images from across the United States. Our methods significantly out-perform the state of the art on two benchmark datasets. We also show, qualitatively, that the proposed feature representations are discriminative at both local and continental spatial scales.

****

Personalized Age Progression With Aging Dictionary

Xiangbo Shu, Jinhui Tang, Hanjiang Lai, Luoqi Liu, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3970-3978

In this paper, we aim to automatically render aging faces in a personalized way. Basically, a set of age-group specific dictionaries are learned, where the dictionary bases corresponding to the same index yet from different dictionaries form a particular aging process pattern cross different age groups, and a linear combination of these patterns expresses a particular personalized aging process. Moreover, two factors are taken into consideration in the dictionary learning process. First, beyond the aging dictionaries, each subject may have extra personalized facial characteristics, e.g. mole, which are invariant in the aging process. Second, it is challenging or even impossible to collect faces of all age groups for a particular subject, yet much easier and more practical to get face pairs from neighboring age groups. Thus a personality-aware coupled reconstruction loss is utilized to learn the dictionaries based on face pairs from neighboring age groups. Extensive experiments well demonstrate the advantages of our proposed solution over other state-of-the-arts in term of personalized aging progression, as well as the performance gain for cross-age face verification by synthesizing aging faces.

****

FaceDirector: Continuous Control of Facial Performance in Video

Charles Malleson, Jean-Charles Bazin, Oliver Wang, Derek Bradley, Thabo Beeler, Adrian Hilton, Alexander Sorkine-Hornung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3979-3987

We present a method to continuously blend between multiple facial performances of an actor, which can contain different facial expressions or emotional states. As an example, given sad and angry video takes of a scene, our method empowers the movie director to specify arbitrary weighted combinations and smooth transitions between the two takes in post-production. Our contributions include (1) a robust nonlinear audio-visual synchronization technique that exploits complementary properties of audio and visual cues to automatically determine robust, dense spatiotemporal correspondences between takes, and (2) a seamless facial blending approach that provides the director full control to interpolate timing, facial expression, and local appearance, in order to generate novel performances after filming. In contrast to most previous works, our approach operates entirely in image space, avoiding the need of 3D facial reconstruction. We demonstrate that our method can synthesize visually believable performances with applications in emotion transition, performance correction, and timing control.

****

Synthesizing Illumination Mosaics From Internet Photo-Collections

Dinghuang Ji, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3988-3996

We propose a framework for the automatic creation of time-lapse mosaics of a given scene. We achieve this by leveraging the illumination variations captured in Internet photo-collections. In order to depict and characterize the illumination spectrum of a scene, our method relies on building discrete representations of the image appearance space through connectivity graphs defined over a pairwise image distance function. The smooth appearance transitions are found as the shortest path in the similarity graph among images, and robust image alignment is ach

ieved by leveraging scene semantics, multi-view geometry, and image warping techniques. The attained results present an insightful and compact visualization of the scene illuminations captured in crowd-sourced imagery.
****

Hot or Not: Exploring Correlations Between Appearance and Temperature
Daniel Glasner, Pascal Fua, Todd Zickler, Lihi Zelnik-Manor; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3997-4005

In this paper we explore interactions between the appearance of an outdoor scene and the ambient temperature. By studying statistical correlations between image sequences from outdoor cameras and temperature measurements we identify two interesting interactions. First, semantically meaningful regions such as foliage and reflective oriented surfaces are often highly indicative of the temperature. Second, small camera motions are correlated with the temperature in some scenes. We propose simple scene-specific temperature prediction algorithms which can be used to turn a camera into a crude temperature sensor. We find that for this task, simple features such as local pixel intensities outperform sophisticated, global features such as from a semantically-trained convolutional neural network.
****

SPM-BP: Sped-up PatchMatch Belief Propagation for Continuous MRFs
Yu Li, Dongbo Min, Michael S. Brown, Minh N. Do, Jiangbo Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4006-4014

Markov random fields are widely used to model many computer vision problems that can be cast in an energy minimization framework composed of unary and pairwise potentials. While computationally tractable discrete optimizers such as Graph Cuts and belief propagation (BP) exist for multi-label discrete problems, they still face prohibitively high computational challenges when the labels reside in a huge or very densely sampled space. Integrating key ideas from PatchMatch of effective particle propagation and resampling, PatchMatch belief propagation (PMBP) has been demonstrated to have good performance in addressing continuous labeling problems and runs orders of magnitude faster than Particle BP (PBP). However, the quality of the PMBP solution is tightly coupled with the local window size, over which the raw data cost is aggregated to mitigate ambiguity in the data constraint. This dependency heavily influences the overall complexity, increasing linearly with the window size. This paper proposes a novel algorithm called sped-up PMBP (SPM-BP) to tackle this critical computational bottleneck and speeds up PMBP by 50-100 times. The crux of SPM-BP is on unifying efficient filter-based cost aggregation and message passing with PatchMatch-based particle generation in a highly effective way. Though simple in its formulation, SPM-BP achieves superior performance for sub-pixel accurate stereo and optical-flow on benchmark datasets when compared with more complex and task-specific approaches.
****

Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation
Christian Bailer, Bertram Taetz, Didier Stricker; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4015-4023

Modern large displacement optical flow algorithms usually use an initialization by either sparse descriptor matching techniques or dense approximate nearest neighbor fields. While the latter have the advantage of being dense, they have the major disadvantage of being very outlier prone as they are not designed to find the optical flow, but the visually most similar correspondence. In this paper we present a dense correspondence field approach that is much less outlier prone and thus much better suited for optical flow estimation than approximate nearest neighbor fields. Our approach is conceptually novel as it does not require explicit regularization, smoothing (like median filtering) or a new data term, but solely our novel purely data based search strategy that finds most inliers (even for small objects), while it effectively avoids finding outliers. Moreover, we present novel enhancements for outlier filtering. We show that our approach is better suited for large displacement optical flow estimation than state-of-the-art descriptor matching techniques. We do so by initializing EpicFlow (so far the best method on MPI-Sintel) with our Flow Fields instead of their originally used s

tate-of-the-art descriptor matching technique. We significantly outperform the original EpicFlow on MPI-Sintel, KITTI and Middlebury.
****

Dense Semantic Correspondence Where Every Pixel is a Classifier
Hilton Bristow, Jack Valmadre, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4024-4031
Determining dense semantic correspondences across objects and scenes is a difficult problem that underpins many higher-level computer vision algorithms. Unlike canonical dense correspondence problems which consider images that are spatially or temporally adjacent, semantic correspondence is characterized by images that share similar high-level structures whose exact appearance and geometry may differ. Motivated by object recognition literature and recent work on rapidly estimating linear classifiers, we treat semantic correspondence as a constrained detection problem, where an exemplar LDA classifier is learned for each pixel. LDA classifiers have two distinct benefits: (i) they exhibit higher average precision than similarity metrics typically used in correspondence problems, and (ii) unlike exemplar SVM, can output globally interpretable posterior probabilities without calibration, whilst also being significantly faster to train. We pose the correspondence problem as a graphical model, where the unary potentials are computed via convolution with the set of exemplar classifiers, and the joint potentials enforce smoothly varying correspondence assignment.
****

Multi-Image Matching via Fast Alternating Minimization
Xiaowei Zhou, Menglong Zhu, Kostas Daniilidis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4032-4040
In this paper we propose a global optimization-based approach to jointly matching a set of images. The estimated correspondences simultaneously maximize pairwise feature affinities and cycle consistency across multiple images. Unlike previous convex methods relying on semidefinite programming, we formulate the problem as a low-rank matrix recovery problem and show that the desired semidefiniteness of a solution can be spontaneously fulfilled. The low-rank formulation enables us to derive a fast alternating minimization algorithm in order to handle practical problems with thousands of features. Both simulation and real experiments demonstrate that the proposed algorithm can achieve a competitive performance with an order of magnitude speedup compared to the state-of-the-art algorithm. In the end, we demonstrate the applicability of the proposed method to match the images of different object instances and as a result the potential to reconstruct category-specific object models from those images.
****

Differential Recurrent Neural Networks for Action Recognition
Vivek Veeriah, Naifan Zhuang, Guo-Jun Qi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4041-4049
The long short-term memory (LSTM) neural network is capable of processing complex sequential information since it utilizes special gating schemes for learning representations from long input sequences. It has the potential to model any time-series or sequential data, where the current hidden state has to be considered in the context of the past hidden states. This property makes LSTM an ideal choice to learn the complex dynamics of various actions. Unfortunately, the conventional LSTMs do not consider the impact of spatio-temporal dynamics corresponding to the given salient motion patterns, when they gate the information that ought to be memorized through time. To address this problem, we propose a differential gating scheme for the LSTM neural network, which emphasizes on the change in information gain caused by the salient motions between the successive frames. This change in information gain is quantified by Derivative of States (DoS), and thus the proposed LSTM model is termed as differential Recurrent Neural Network (dRNN). We demonstrate the effectiveness of the proposed model by automatically recognizing actions from the real-world 2D and 3D human action datasets. Our study is one of the first works towards demonstrating the potential of learning complex time-series representations via high-order derivatives of states.
****

Similarity Gaussian Process Latent Variable Model for Multi-Modal Data Analysis

Guoli Song, Shuhui Wang, Qingming Huang, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4050-4058

Data from real applications involve multiple modalities representing content with the same semantics and deliver rich information from complementary aspects. However, relations among heterogeneous modalities are simply treated as observation-to-fit by existing work, and the parameterized cross-modal mapping functions lack flexibility in directly adapting to the content divergence and semantic complicacy of multi-modal data. In this paper, we build our work based on Gaussian process latent variable model (GPLVM) to learn the non-linear non-parametric mapping functions and transform heterogeneous data into a shared latent space. We propose multi-modal Similarity Gaussian Process latent variable model (m-SimGP), which learns the nonlinear mapping functions between the intra-modal similarities and latent representation. We further propose multi-modal regularized similarity GPLVM (m-RSimGP) by encouraging similar/dissimilar points to be similar/dissimilar in the output space. The overall objective functions are solved by simple and scalable gradient decent techniques. The proposed models are robust to content divergence and high-dimensionality in multi-modal representation. They can be applied to various tasks to discover the non-linear correlations and obtain the comparable low-dimensional representation for heterogeneous modalities. On two widely used real-world datasets, we outperform previous approaches for cross-modal content retrieval and cross-modal classification.

****

Learning Ensembles of Potential Functions for Structured Prediction With Latent Variables

Hossein Hajimirsadeghi, Greg Mori; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4059-4067

Many visual recognition tasks involve modeling variables which are structurally related. Hidden conditional random fields (HCRFs) are a powerful class of models for encoding structure in weakly supervised training examples. This paper presents HCRF-Boost, a novel and general framework for learning HCRFs in functional space. An algorithm is proposed to learn the potential functions of an HCRF as a combination of abstract nonlinear feature functions, expressed by regression models. Consequently, the resulting latent structured model is not restricted to traditional log-linear potential functions or any explicit parameterization. Further, functional optimization helps to avoid direct interactions with the possibly large parameter space of nonlinear models and improves efficiency. As a result, a complex and flexible ensemble method is achieved for structured prediction which can be successfully used in a variety of applications. We validate the effectiveness of this method on tasks such as group activity recognition, human action recognition, and multi-instance learning of video events.

****

Simultaneous Deep Transfer Across Domains and Tasks

Eric Tzeng, Judy Hoffman, Trevor Darrell, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4068-4076

Recent reports suggest that a generic supervised deep CNN model trained on a large-scale dataset reduces, but does not remove, dataset bias. Fine-tuning deep models in a new domain can require a significant amount of labeled data, which for many applications is simply not available. We propose a new CNN architecture to exploit unlabeled and sparsely labeled target domain data. Our approach simultaneously optimizes for domain invariance to facilitate domain transfer and uses a soft label distribution matching loss to transfer information between tasks. Our proposed adaptation method offers empirical performance which exceeds previously published results on two standard benchmark visual domain adaptation tasks, evaluated across supervised and semi-supervised adaptation settings.

****

Low Dimensional Explicit Feature Maps

Ondrej Chum; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4077-4085

Approximating non-linear kernels by finite-dimensional feature maps is a popular approach for speeding up training and evaluation of support vector machines or to encode information into efficient match kernels. We propose a novel method of data independent construction of low dimensional feature maps. The problem is cast as a linear program which jointly considers competing objectives: the quality of the approximation and the dimensionality of the feature map. For both shift-invariant and homogeneous kernels the proposed method achieves a better approximations at the same dimensionality or comparable approximations at lower dimensionality of the feature map compared with state-of-the-art methods.

****

Unsupervised Learning of Spatiotemporally Coherent Metrics

Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, Yann LeCun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4086-4093

Current state-of-the-art classification and detection algorithms train deep convolutional networks using labeled data. In this work we study unsupervised feature learning with convolutional networks in the context of temporally coherent unlabeled data. We focus on feature learning from unlabeled video data, using the assumption that adjacent video frames contain semantically similar information. This assumption is exploited to train a convolutional pooling auto-encoder regularized by slowness and sparsity. We establish a connection between slow feature learning and metric learning. Using this connection we define "temporal coherence"--a criterion which can be used to select hyper-parameters automatically. In a transfer learning experiment, we show that the resulting encoder can be used to define a more semantically coherent metric without the use of labeled data.

****

Multi-Label Cross-Modal Retrieval

Viresh Ranjan, Nikhil Rasiwasia, C. V. Jawahar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4094-4102

In this work, we address the problem of cross-modal retrieval in presence of multi-label annotations. In particular, we introduce multi-label Canonical Correlation Analysis (ml-CCA), an extension of CCA, for learning shared subspaces taking into account high level semantic information in the form of multi-label annotations. Unlike CCA, ml-CCA does not rely on explicit pairing between modalities, instead it uses the multi-label information to establish correspondences. This results in a discriminative subspace which is better suited for cross-modal retrieval tasks. We also present Fast ml-CCA, a computationally efficient version of ml-CCA, which is able to handle large scale datasets. We show the efficacy of our approach by conducting extensive cross-modal retrieval experiments on three standard benchmark datasets. The results show that the proposed approach achieves state of the art retrieval performance on the three datasets.

****

Improving Ferns Ensembles by Sparsifying and Quantising Posterior Probabilities

Antonio L. Rodriguez, Vitor Sequeira; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4103-4111

Ferns ensembles offer an accurate and efficient multiclass non-linear classification, commonly at the expense of consuming a large amount of memory. We introduce a two-fold contribution that produces large reductions in their memory consumption. First, an efficient L0 regularised cost optimisation finds a sparse representation of the posterior probabilities in the ensemble by discarding elements with zero contribution to valid responses in the training samples. As a by-product this can produce a prediction accuracy gain that, if required, can be traded for further reductions in memory size and prediction time. Secondly, posterior probabilities are quantised and stored in a memory-friendly sparse data structure. We reported a minimum of 75% memory reduction for different types of classification problems using generative and discriminative ferns ensembles, without increasing prediction time or classification error. For image patch recognition our proposal produced a 90% memory reduction, and improved in several percentage points the prediction accuracy.

****

Beyond Gauss: Image-Set Matching on the Riemannian Manifold of PDFs

Mehrtash Harandi, Mathieu Salzmann, Mahsa Baktashmotlagh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4112-4120

State-of-the-art image-set matching techniques typically implicitly model each image-set with a Gaussian distribution. Here, we propose to go beyond these representations and model image-sets as probability distribution functions (PDFs) using kernel density estimators. To compare and match image-sets, we exploit Csiszar f-divergences, which bear strong connections to the geodesic distance defined on the space of PDFs, i.e., the statistical manifold. Furthermore, we introduce valid positive definite kernels on the statistical manifolds, which let us make use of more powerful classification schemes to match image-sets. Finally, we introduce a supervised dimensionality reduction technique that learns a latent space where f-divergences reflect the class labels of the data. Our experiments on diverse problems, such as video-based face recognition and dynamic texture classification, evidence the benefits of our approach over the state-of-the-art image-set matching methods.

****

Unsupervised Domain Adaptation With Imbalanced Cross-Domain Data

Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, Yu-Chiang Frank Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4121-4129

We address a challenging unsupervised domain adaptation problem with imbalanced cross-domain data. For standard unsupervised domain adaptation, one typically obtains labeled data in the source domain and only observes unlabeled data in the target domain. However, most existing works do not consider the scenarios in which either the label numbers across domains are different, or the data in the source and/or target domains might be collected from multiple datasets. To address the aforementioned settings of imbalanced cross-domain data, we propose Closest Common Space Learning (CCSL) for associating such data with the capability of preserving label and structural information within and across domains. Experiments on multiple cross-domain visual classification tasks confirm that our method performs favorably against state-of-the-art approaches, especially when imbalanced cross-domain data are presented.

****

Secrets of Matrix Factorization: Approximations, Numerics, Manifold Optimization and Random Restarts

Je Hyeong Hong, Andrew Fitzgibbon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4130-4138

Matrix factorization (or low-rank matrix completion) with missing data is a key computation in many computer vision and machine learning tasks, and is also related to a broader class of nonlinear optimization problems such as bundle adjustment. The problem has received much attention recently, with renewed interest in variable-projection approaches, yielding dramatic improvements in reliability and speed. However, on a wide class of problems, no one approach dominates, and because the various approaches have been derived in a multitude of different ways, it has been difficult to unify them. This paper provides a unified derivation of a number of recent approaches, so that similarities and differences are easily observed. We also present a simple meta-algorithm which wraps any existing algorithm, yielding 100% success rate on many standard datasets. Given 100% success, the focus of evaluation must turn to speed, as 100% success is trivially achieved if we do not care about speed. Again our unification allows a number of generic improvements applicable to all members of the family to be isolated, yielding a unified algorithm that outperforms our re-implementation of existing algorithms, which in some cases already outperform the original authors' publicly available codes.

****

Geometry-Aware Deep Transform

Jiaji Huang, Qiang Qiu, Robert Calderbank, Guillermo Sapiro; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4139-4147

Many recent efforts have been devoted to designing sophisticated deep learning structures, obtaining revolutionary results on benchmark datasets. The success of these deep learning methods mostly relies on an enormous volume of labeled training samples to learn a huge number of parameters in a network; therefore, understanding the generalization ability of a learned deep network cannot be overlooked, especially when restricted to a small training set, which is the case for many applications. In this paper, we propose a novel deep learning objective formulation that unifies both the classification and metric learning criteria. We then introduce a geometry-aware deep transform to enable a non-linear discriminative and robust feature transform, which shows competitive performance on small training sets for both synthetic and real-world data. We further support the proposed framework with a formal (K,epsilon)-robustness analysis.

****

Learning Binary Codes for Maximum Inner Product Search

Fumin Shen, Wei Liu, Shaoting Zhang, Yang Yang, Heng Tao Shen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4148-4156

Binary coding or hashing techniques are recognized to accomplish efficient near neighbor search, and have thus attracted broad interests in the recent vision and learning studies. However, such studies have rarely been dedicated to Maximum Inner Product Search (MIPS), which plays a critical role in various vision applications. In this paper, we investigate learning binary codes to exclusively handle the MIPS problem. Inspired by the latest advance in asymmetric hashing schemes, we propose an asymmetric binary code learning framework based on inner product fitting. Specifically, two sets of coding functions are learned such that the inner products between their generated binary codes can reveal the inner products between original data vectors. We also propose an alternative simpler objective which maximizes the correlations between the inner products of the produced binary codes and raw data vectors. In both objectives, the binary codes and coding functions are simultaneously learned without continuous relaxations, which is the key to achieving high-quality binary codes. We evaluate the proposed method, dubbed Asymmetric Inner-product Binary Coding (AIBC), relying on the two objectives on several large-scale image datasets. Both of them are superior to the state-of-the-art binary coding and hashing methods in performing MIPS tasks.

****

ML-MG: Multi-Label Learning With Missing Labels Using a Mixed Graph

Baoyuan Wu, Siwei Lyu, Bernard Ghanem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4157-4165

This work focuses on the problem of multi-label learning with missing labels (MLML), which aims to label each test instance with multiple class labels given training instances that have an incomplete/partial set of these labels (i.e. some of their labels are missing). To handle missing labels, we propose a unified model of label dependencies by constructing a mixed graph, which jointly incorporates (i) instance-level similarity and class co-occurrence as undirected edges and (ii) semantic label hierarchy as directed edges. Unlike most MLML methods, We formulate this learning problem transductively as a convex quadratic matrix optimization problem that encourages training label consistency and encodes both types of label dependencies (i.e. undirected and directed edges) using quadratic terms and hard linear constraints. The alternating direction method of multipliers (ADMM) can be used to exactly and efficiently solve this problem. To evaluate our proposed method, we consider two popular applications (image and video annotation), where the label hierarchy can be derived from Wordnet. Experimental results show that our method achieves a significant improvement over state-of-the-art methods in performance and robustness to missing labels.

****

Zero-Shot Learning via Semantic Similarity Embedding

Ziming Zhang, Venkatesh Saligrama; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4166-4174

In this paper we consider a version of the zero-shot learning problem where seen class source and target domain data are provided. The goal during test-time is to accurately predict the class label of an unseen target domain instance based on revealed source domain side information (e.g. attributes) for unseen classes. Our method is based on viewing each source or target data as a mixture of seen class proportions and we postulate that the mixture patterns have to be similar if the two instances belong to the same unseen class. This perspective leads us to learning source/target embedding functions that map an arbitrary source/target domain data into a same semantic space where similarity can be readily measured. We develop a max-margin framework to learn these similarity functions and jointly optimize parameters by means of cross validation. Our test results are compelling, leading to significant improvement in terms of accuracy on most benchmark datasets for zero-shot recognition.

****

Bayesian Model Adaptation for Crowd Counts
Bo Liu, Nuno Vasconcelos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4175-4183

The problem of transfer learning is considered in the domain of crowd counting. A solution based on Bayesian model adaptation of Gaussian processes is proposed. This is shown to produce intuitive model updates, which are tractable, and lead to an adapted model (predictive distribution) that accounts for all information in both training and adaptation data. The new adaptation procedure achieves significant gains over previous approaches, based on multi-task learning, while requiring much less computation to deploy. This makes it particularly suited for the problem of expanding the capacity of crowd counting camera networks. A large video dataset for the evaluation of adaptation approaches to crowd counting is also introduced. This contains a number of adaptation tasks, involving information transfer across video collected by 1) a single camera under different scene conditions (different times of the day) and 2) video collected from different cameras. Evaluation of the proposed model adaptation procedure in this dataset shows good performance in realistic operating conditions.

****

An NMF Perspective on Binary Hashing
Lopamudra Mukherjee, Sathya N. Ravi, Vamsi K. Ithapu, Tyler Holmes, Vikas Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4184-4192

The pervasiveness of massive data repositories has led to much interest in efficient methods for indexing, search, and retrieval. For image data, a rapidly developing body of work for these applications shows impressive performance with methods that broadly fall under the umbrella term of Binary Hashing. Given a distance matrix, a binary hashing algorithm solves for a binary code for the given set of examples, whose Hamming distance nicely approximates the original distances. The formulation is non-convex-- so existing solutions adopt spectral relaxations or perform coordinate descent (or quantization) on a surrogate objective that is numerically more tractable. In this paper, we first derive an Augmented Lagrangian approach to optimize the standard binary Hashing objective (i.e.,maintain fidelity with a given distance matrix). With appropriate step sizes, we find that this scheme already yields results that match or substantially outperform state of the art methods on most benchmarks used in the literature. Then, to allow the model to scale to large datasets, we obtain an interesting reformulation of the binary hashing objective as a non negative matrix factorization. Later, this leads to a simple multiplicative updates algorithm -- whose parallelization properties are exploited to obtain a fast GPU based implementation. We give a probabilistic analysis of our initialization scheme and present a range of experiments to show that the method is simple to implement and competes favorably with available methods (both for optimization and generalization).

****

Multi-View Domain Generalization for Visual Recognition
Li Niu, Wen Li, Dong Xu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4193-4201

In this paper, we propose a new multi-view domain generalization (MVDG) approach for visual recognition, in which we aim to use the source domain samples with multiple types of features (i.e., multi-view features) to learn robust classifiers that can generalize well to any unseen target domain. Considering the recent works show the domain generalization capability can be enhanced by fusing multipl e SVM classifiers, we build upon exemplar SVMs to learn a set of SVM classifiers by using one positive sample and all negative samples in the source domain each time. When the source domain samples come from multiple latent domains, we expect the weight vectors of exemplar SVM classifiers can be organized into multiple hidden clusters. To exploit such cluster structure, we organize the weight vectors learnt on each view as a weight matrix and seek the low-rank representation by reconstructing this weight matrix using itself as the dictionary. To enforce the consistency of inherent cluster structures discovered from the weight matrices learnt on different views, we introduce a new regularizer to minimize the mismatch between any two representation matrices on different views. We also develop an efficient alternating optimization algorithm and further extend our MVDG approach for domain adaptation by exploiting the manifold structure of unlabeled target domain samples. Comprehensive experiments for visual recognition clearly demonstrate the effectiveness of our approaches for domain generalization and domain adaptation.

****

Infinite Feature Selection

Giorgio Roffo, Simone Melzi, Marco Cristani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4202-4210

Filter-based feature selection has become crucial in many classification settings, especially object recognition, recently faced with feature learning strategies that originate thousands of cues. In this paper, we propose a feature selection method exploiting the convergence properties of power series of matrices, and introducing the concept of infinite feature selection (Inf-FS). Considering a selection of features as a path among feature distributions and letting these paths tend to an infinite number permits the investigation of the importance (relevance and redundancy) of a feature when injected into an arbitrary set of cues. Ranking the importance individuates candidate features, which turn out to be effective from a classification point of view, as proved by a thoroughly experimental section. The Inf-FS has been tested on thirteen diverse benchmarks, comparing against filters, embedded methods, and wrappers; in all the cases we achieve top performances, notably on the classification tasks of PASCAL VOC 2007-2012.

****

Semi-Supervised Zero-Shot Classification With Label Representation Learning

Xin Li, Yuhong Guo, Dale Schuurmans; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4211-4219

Given the challenge of gathering labeled training data, zero-shot classification, which transfers information from observed classes to recognize unseen classes, has become increasingly popular in the computer vision community. Most existing zero-shot learning methods require a user to first provide a set of semantic visual attributes for each class as side information before applying a two-step prediction procedure that introduces an intermediate attribute prediction problem. In this paper, we propose a novel zero-shot classification approach that automatically learns label embeddings from the input data in a semi-supervised large-margin learning framework. The proposed framework jointly considers multi-class classification over all classes (observed and unseen) and tackles the target prediction problem directly without introducing intermediate prediction problems. It also has the capacity to incorporate semantic label information from different sources when available. To evaluate the proposed approach, we conduct experiments on standard zero-shot data sets. The empirical results show the proposed approach outperforms existing state-of-the-art zero-shot learning methods.

****

A Supervised Low-Rank Method for Learning Invariant Subspaces

Farzad Siyahjani, Ranya Almohsen, Sinan Sabri, Gianfranco Doretto; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4220-4228

Sparse representation and low-rank matrix decomposition approaches have been successfully applied to several computer vision problems. They build a generative representation of the data, which often requires complex training as well as testing to be robust against data variations induced by nuisance factors. We introdu ce the invariant components, a discriminative representation invariant to nuisance factors, because it spans subspaces orthogonal to the space where nuisance factors are defined. This allows developing a framework based on geometry that ensures a uniform inter-class separation, and a very efficient and robust classification based on simple nearest neighbor. In addition, we show how the approach is equivalent to a local metric learning, where the local metrics (one for each class) are learned jointly, rather than independently, thus avoiding the risk of overfitting without the need for additional regularization. We evaluated the approach for face recognition with highly corrupted training and testing data, obtaining very promising results.

****

Recursive Frechet Mean Computation on the Grassmannian and its Applications to Computer Vision

Rudrasis Chakraborty, Baba C. Vemuri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4229-4237

In the past decade, Grassmann manifolds (Grassmannian) have been commonly used in mathematical formulations of many Computer Vision tasks. Averaging points on a Grassmann manifold is a very common operation in many applications including but not limited to, tracking, action recognition, video-face recognition, face recognition, etc. Computing the intrinsic/Frechet mean (FM) of a set of points on the Grassmann can be cast as finding the global optimum (if it exists) of the sum of squared geodesic distances cost function. A common approach to solve this problem involves the use of the gradient descent method. An alternative way to compute the FM is to develop a recursive/inductive definition that does not involve optimizing the aforementioned cost function. In this paper, we propose one such computationally efficient algorithm called the it Grassmann   inductive Frechet mean estimator (GiFME). In developing the recursive solution to find the FM of the given set of points, GiFME exploits the fact that there is a closed form solution to find the FM of two points on the Grassmann. In the limit as the number of samples tends to infinity, we prove that GiFME converges to the FM (this is called the weak consistency result on the Grassmann manifold). Further, for the finite sample case, in the limit as the number of sample paths (trials) goes to infinity, we show that GiFME converges to the finite sample FM. Moreover, we present a bound on the geodesic distance between the estimate from GiFME and the true FM. We present several experiments on synthetic and real data sets to demonstrate the performance of GiFME in comparison to the gradient descent based (batch mode) technique. Our goal in these applications is to demonstrate the computational advantage and achieve comparable accuracy to the state-of-the-art.

****

Multi-View Subspace Clustering

Hongchang Gao, Feiping Nie, Xuelong Li, Heng Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4238-4246

For many computer vision applications, the data sets distribute on certain low-dimensional subspaces. Subspace clustering is to find such underlying subspaces and cluster the data points correctly. In this paper, we propose a novel multi-view subspace clustering method. The proposed method performs clustering on the subspace representation of each view simultaneously. Meanwhile, we propose to use a common cluster structure to guarantee the consistence among different views. In addition, an efficient algorithm is proposed to solve the problem. Experiments on four benchmark data sets have been performed to validate our proposed method. The promising results demonstrate the effectiveness of our method.

****

Predicting Deep Zero-Shot Convolutional Neural Networks Using Textual Descriptions

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, Ruslan salakhutdinov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4247-4255

One of the main challenges in Zero-Shot Learning of visual categories is gathering semantic attributes to accompany images. Recent work has shown that learning from textual descriptions, such as Wikipedia articles, avoids the problem of having to explicitly define these attributes. We present a new model that can class ify unseen categories from their textual description. Specifically, we use text features to predict the output weights of both the convolutional and the fully connected layers in a deep convolutional neural network (CNN). We take advantage of the architecture of CNNs and learn features at different layers, rather than just learning an embedding space for both modalities, as is common with existing approaches. The proposed model also allows us to automatically generate a list of pseudo-attributes for each visual category consisting of words from Wikipedia articles. We train our models end-to-end using the Caltech-UCSD bird and flower datasets and evaluate both ROC and Precision-Recall curves. Our empirical results show that the proposed model significantly outperforms previous methods.

****

Structured Feature Selection

Tian Gao, Ziheng Wang, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4256-4264

Feature dimensionality reduction has been widely used in various computer vision tasks. We explore feature selection as the dimensionality reduction technique and propose to use a structured approach, based on the Markov Blanket (MB), to select features. We first introduce a new MB discovery algorithm, Simultaneous Markov Blanket (STMB) discovery, that improves the efficiency of state-of-the-art algorithms. Then we theoretically justify three advantages of structured feature selection over traditional feature selection methods. Specifically, we show that the Markov Blanket is the minimum feature set that retains the maximal mutual information and also gives the lowest Bayes classification error. Then we apply structured feature selection to two applications: 1) We introduce a new method that enables STMB to scale up and show the competitive performance of our algorithms on large-scale image classification tasks. 2) We propose a method for structured feature selection to handle hierarchical features and show the proposed method can lead to big performance gain in facial expression and action unit (AU) recognition tasks.

****

Conditional High-Order Boltzmann Machine: A Supervised Learning Model for Relation Learning

Yan Huang, Wei Wang, Liang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4265-4273

Relation learning is a fundamental operation in many computer vision tasks. Recently, high-order Boltzmann machine and its variants have exhibited the great power of modelling various data relation. However, most of them are unsupervised learning models which are not very discriminative and thus cannot server as a standalone solution to relation learning tasks. In this paper, we explore supervised learning algorithms and propose a new model named Conditional High-order Boltzmann Machine (CHBM), which can be directly used as a bilinear classifier to assign similarity scores for pairwise images. Then, to better deal with complex data relation, we propose a gated version of CHBM which untangles factors of variation by exploiting a set of latent variables to gate classification. We perform four-order tensor factorization for parameter reduction, and present two efficient supervised learning algorithms from the perspectives of being generative and discriminative, respectively. The experimental results of image transformation visualization, binary-way classification and face verification demonstrate that, by performing supervised learning, our models can greatly improve the performance.

****

Learning Image and User Features for Recommendation in Social Networks

Xue Geng, Hanwang Zhang, Jingwen Bian, Tat-Seng Chua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4274-4282

Good representations of data do help in many machine learning tasks such as recommendation. It is often a great challenge for traditional recommender systems to learn representative features of both users and images in large social networks, in particular, social curation networks, which are characterized as the extremely sparse links between users and images, and the extremely diverse visual contents of images. To address the challenges, we propose a novel deep model which learns the unified feature representations for both users and images. This is don

e by transforming the heterogeneous user-image networks into homogeneous low-dimensional representations, which facilitate a recommender to trivially recommend images to users by feature similarity. We also develop a fast online algorithm that can be easily scaled up to large networks in an asynchronously parallel way. We conduct extensive experiments on a representative subset of Pinterest, containing 1,456,540 images and 1,000,000 users. Results of image recommendation experiments demonstrate that our feature learning approach significantly outperforms other state-of-the-art recommendation methods.
****
Dual-Feature Warping-Based Motion Model Estimation
Shiwei Li, Lu Yuan, Jian Sun, Long Quan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4283-4291
To break down the geometry assumptions of traditional motion models (e.g., homography, affine), warping-based motion model recently becomes very popular and is adopted in many latest applications (e.g., image stitching, video stabilization). With high degrees of freedom, the accuracy of model heavily relies on data-terms (keypoint correspondences). In some low-texture environments (e.g., indoor) where keypoint feature is insufficient or unreliable, the warping model is often erroneously estimated. In this paper we propose a simple and effective approach by considering both keypoint and line segment correspondences as data-term. Line segment is a prominent feature in artificial environments and it can supply sufficient geometrical and structural information of scenes, which not only helps guild to a correct warp in low-texture condition, but also prevents the undesired distortion induced by warping. The combination aims to complement each other and benefit for a wider range of scenes. Our method is general and can be ported to many existing applications. Experiments demonstrate that using dual-feature yields more robust and accurate result especially for those low-texture images.
****
An Adaptive Data Representation for Robust Point-Set Registration and Merging
Dylan Campbell, Lars Petersson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4292-4300
This paper presents a framework for rigid point-set registration and merging using a robust continuous data representation. Our point-set representation is constructed by training a one-class support vector machine with a Gaussian radial basis function kernel and subsequently approximating the output function with a Gaussian mixture model. We leverage the representation's sparse parametrisation and robustness to noise, outliers and occlusions in an efficient registration algorithm that minimises the L2 distance between our support vector-parametrised Gaussian mixtures. In contrast, existing techniques, such as Iterative Closest Point and Gaussian mixture approaches, manifest a narrower region of convergence and are less robust to occlusions and missing data, as demonstrated in the evaluation on a range of 2D and 3D datasets. Finally, we present a novel algorithm, GMMerge, that parsimoniously and equitably merges aligned mixture models, allowing the framework to be used for reconstruction and mapping.
****
Local Subspace Collaborative Tracking
Lin Ma, Xiaoqin Zhang, Weiming Hu, Junliang Xing, Jiwen Lu, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4301-4309
Subspace models have been widely used for appearance based object tracking. Most existing subspace based trackers employ a linear subspace to represent object appearances, which are not accurate enough to model large variations of objects. To address this, this paper presents a local subspace collaborative tracking method for robust visual tracking, where multiple linear and nonlinear subspaces are learned to better model the nonlinear relationship of object appearances. First, we retain a set of key samples and compute a set of local subspaces for each key sample. Then, we construct a hyper sphere to represent the local nonlinear subspace for each key sample. The hyper sphere of one key sample passes the local key samples and also is tangent to the local linear subspace of the specific key sample. In this way, we are able to represent the nonlinear distribution of th

e key samples and also approximate the local linear subspace near the specific key sample, so that local distributions of the samples can be represented more accurately. Experimental results on challenging video sequences demonstrate the effectiveness of our method.

****

Learning Spatially Regularized Correlation Filters for Visual Tracking

Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, Michael Felsberg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4310-4318

Robust and accurate visual tracking is one of the most challenging computer vision problems. Due to the inherent lack of training data, a robust approach for constructing a target appearance model is crucial. Recently, discriminatively learned correlation filters (DCF) have been successfully applied to address this problem for tracking. These methods utilize a periodic assumption of the training samples to efficiently learn a classifier on all patches in the target neighborhood. However, the periodic assumption also introduces unwanted boundary effects, which severely degrade the quality of the tracking model. We propose Spatially Regularized Discriminative Correlation Filters (SRDCF) for tracking. A spatial regularization component is introduced in the learning to penalize correlation filter coefficients depending on their spatial location. Our SRDCF formulation allows the correlation filters to be learned on a significantly larger set of negative training samples, without corrupting the positive samples. We further propose an optimization strategy, based on the iterative Gauss-Seidel method, for efficient online learning of our SRDCF. Experiments are performed on four benchmark datasets: OTB-2013, ALOV++, OTB-2015, and VOT2014. Our approach achieves state-of-the-art results on all four datasets. On OTB-2013 and OTB-2015, we obtain an absolute gain of 8.0% and 8.2% respectively, in mean overlap precision, compared to the best existing trackers.

****

SpeDo: 6 DOF Ego-Motion Sensor Using Speckle Defocus Imaging

Kensei Jo, Mohit Gupta, Shree K. Nayar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4319-4327

Sensors that measure their motion with respect to the surrounding environment (ego-motion sensors) can be broadly classified into two categories. First is inertial sensors such as accelerometers. In order to estimate position and velocity, these sensors integrate the measured acceleration, which often results in accumulation of large errors over time. Second, camera-based approaches such as SLAM that can measure position directly, but their performance depends on the surrounding scene's properties. These approaches cannot function reliably if the scene has low frequency textures or small depth variations. We present a novel ego-motion sensor called SpeDo that addresses these fundamental limitations. SpeDo is based on using coherent light sources and cameras with large defocus. Coherent light, on interacting with a scene, creates a high frequency interferometric pattern in the captured images, called speckle. We develop a theoretical model for speckle flow (motion of speckle as a function of sensor motion), and show that it is quasi-invariant to surrounding scene's properties. As a result, SpeDo can measure ego-motion (not derivative of motion) simply by estimating optical flow at a few image locations. We have built a low-cost and compact hardware prototype of SpeDo and demonstrated high precision 6 DOF ego-motion estimation for complex trajectories in scenarios where the scene properties are challenging (e.g., repeating or no texture) as well as unknown.

****

Unsupervised Trajectory Clustering via Adaptive Multi-Kernel-Based Shrinkage

Hongteng Xu, Yang Zhou, Weiyao Lin, Hongyuan Zha; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4328-4336

This paper proposes a shrinkage-based framework for unsupervised trajectory clustering. Facing to the challenges of trajectory clustering, e.g., large variations within a cluster and ambiguities across clusters, we first introduce an adaptive multi-kernel-based estimation process to estimate the `shrunk' positions and speeds of trajectories' points. This kernel-based estimation effectively leverag es both multiple structural information within a trajectory and the local motion patterns across multiple trajectories, such that the discrimination of the shrunk point can be properly increased. We further introduce a speed-regularized optimization process, which utilizes the estimated speeds to regularize the optimal shrunk points, so as to guarantee the smoothness and the discriminative pattern of the final shrunk trajectory. Using our approach, the variations among similar trajectories can be reduced while the boundaries between different clusters are enlarged. Experimental results demonstrate that our approach is superior to the state-of-art approaches on both clustering accuracy and robustness. Besides, additional experiments further reveal the effectiveness of our approach when applied to trajectory analysis applications such as anomaly detection.

****

TRIC-track: Tracking by Regression With Incrementally Learned Cascades

Xiaomeng Wang, Michel Valstar, Brais Martinez, Muhammad Haris Khan, Tony Pridmore; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4337-4345

This paper proposes a novel approach to part-based tracking by replacing local matching of an appearance model by direct prediction of the displacement between local image patches and part locations. We propose to use cascaded regression with incremental learning to track generic objects without any prior knowledge of an object's structure or appearance. We exploit the spatial constraints between parts by implicitly learning the shape and deformation parameters of the object in an online fashion. We integrate a multiple temporal scale motion model to initialise our cascaded regression search close to the target and to allow it to cope with occlusions. Experimental results show that our tracker ranks first on the CVPR 2013 Benchmark.

****

Recurrent Network Models for Human Dynamics

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4346-4354

We propose the Encoder-Recurrent-Decoder (ERD) model for recognition and prediction of human body pose in videos and motion capture. The ERD model is a recurrent neural network that incorporates nonlinear encoder and decoder networks before and after recurrent layers. We test instantiations of ERD architectures in the tasks of motion capture (mocap) generation, body pose labeling and body pose forecasting in videos. Our model handles mocap training data across multiple subjects and activity domains, and synthesizes novel motions while avoiding drifting for long periods of time. For human pose labeling, ERD outperforms a per frame body part detector by resolving left-right body part confusions. For video pose forecasting, ERD predicts body joint displacements across a temporal horizon of 400ms and outperforms a first order motion model based on optical flow. ERDs extend previous Long Short Term Memory (LSTM) models in the literature to jointly learn representations and their dynamics. Our experiments show such representation learning is crucial for both labeling and prediction in space-time. We find this is a distinguishing feature between the spatio-temporal visual domain in comparison to 1D text, speech or handwriting, where straightforward hard coded representations have shown excellent results when directly combined with recurrent units.

****

Contour Flow: Middle-Level Motion Estimation by Combining Motion Segmentation and Contour Alignment

Huijun Di, Qingxuan Shi, Feng Lv, Ming Qin, Yao Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4355-4363

Our goal is to estimate contour flow (the contour pairs with consistent point correspondence) from inconsistent contours extracted independently in two video frames. We formulate the contour flow estimation locally as a motion segmentation problem where motion patterns grouped from optical flow field are exploited for local correspondence measurement. To solve local ambiguities, contour flow estimation is further formulated globally as a contour alignment problem. We propose a novel two-staged strategy to obtain global consistent point correspondence under various contour transitions such as splitting, merging and branching. The goal of the first stage is to obtain possible accurate contour-to-contour alignments, and the second stage aims to make a consistent fusion of many partial alignments. Such a strategy can properly balance the accuracy and the consistency, which enables a middle-level motion representation to be constructed by just concatenating frame-by-frame contour flow estimation. Experiments prove the effectiveness of our method.

****

FollowMe: Efficient Online Min-Cost Flow Tracking With Bounded Memory and Computation

Philip Lenz, Andreas Geiger, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4364-4372

One of the most popular approaches to multi-target tracking is tracking-by-detection. Current min-cost flow algorithms which solve the data association problem optimally have three main drawbacks: they are computationally expensive, they assume that the whole video is given as a batch, and they scale badly in memory and computation with the length of the video sequence. In this paper, we address each of these issues, resulting in a computationally and memory-bounded solution. First, we introduce a dynamic version of the successive shortest-path algorithm which solves the data association problem optimally while reusing computation, resulting in faster inference than standard solvers. Second, we address the optimal solution to the data association problem when dealing with an incoming stream of data (i.e., online setting). Finally, we present our main contribution which is an approximate online solution with bounded memory and computation which is capable of handling videos of arbitrary length while performing tracking in real time. We demonstrate the effectiveness of our algorithms on the KITTI and PETS2009 benchmarks and show state-of-the-art performance, while being significantly faster than existing solvers.

****

Learning to Divide and Conquer for Online Multi-Target Tracking

Francesco Solera, Simone Calderara, Rita Cucchiara; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4373-4381

Online Multiple Target Tracking (MTT) is often addressed within the tracking-by-detection paradigm. Detections are previously extracted independently in each frame and then objects trajectories are built by maximizing specifically designed coherence functions. Nevertheless, ambiguities arise in presence of occlusions or detection errors. In this paper we claim that the ambiguities in tracking could be solved by a selective use of the features, by working with more reliable features if possible and exploiting a deeper representation of the target only if necessary. To this end, we propose an online divide and conquer tracker for static camera scenes, which partitions the assignment problem in local subproblems and solves them by selectively choosing and combining the best features. The complete framework is cast as a structural learning task that unifies these phases and learns tracker parameters from examples. Experiments on two different datasets highlights a significant improvement of tracking performances (MOTA +10%) over the state of the art.

****

Minimizing Human Effort in Interactive Tracking by Incremental Learning of Model Parameters

Arridhana Ciptadi, James M. Rehg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4382-4390

We address the problem of minimizing human effort in interactive tracking by learning sequence-specific model parameters. Determining the optimal model parameters for each sequence is a critical problem in tracking. We demonstrate that by using the optimal model parameters for each sequence we can achieve high precision tracking results with significantly less effort. We leverage the sequential nature of interactive tracking to formulate an efficient method for learning model parameters through a maximum margin framework. By using our method we are able to save 60-90% of human effort to achieve high precision on two datasets: the VI

RAT dataset and an Infant-Mother Interaction dataset.

****

A Novel Representation of Parts for Accurate 3D Object Detection and Tracking in Monocular Images

Alberto Crivellaro, Mahdi Rad, Yannick Verdie, Kwang Moo Yi, Pascal Fua, Vincent Lepetit; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4391-4399

We present a method that estimates in real-time and under challenging conditions the 3D pose of a known object. Our method relies only on grayscale images since depth cameras fail on metallic objects; it can handle poorly textured objects, and cluttered, changing environments; the pose it predicts degrades gracefully in presence of large occlusions. As a result, by contrast with the state-of-the-art, our method is suitable for practical Augmented Reality applications even in industrial environments. To be robust to occlusions, we first learn to detect some parts of the target object. Our key idea is to then predict the 3D pose of each part in the form of the 2D projections of a few control points. The advantages of this representation is three-fold: We can predict the 3D pose of the object even when only one part is visible; when several parts are visible, we can combine them easily to compute a better pose of the object; the 3D pose we obtain is usually very accurate, even when only few parts are visible.

****

Linearization to Nonlinear Learning for Visual Tracking

Bo Ma, Hongwei Hu, Jianbing Shen, Yuping Zhang, Fatih Porikli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4400-4407

Due to unavoidable appearance variations caused by occlusion, deformation, and other factors, classifiers for visual tracking are nonlinear as a necessity. Building on the theory of globally linear approximations to nonlinear functions, we introduce an elegant method that jointly learns a nonlinear classifier and a visual dictionary for tracking objects in a semi-supervised sparse coding fashion. This establishes an obvious distinction from conventional sparse coding based discriminative tracking algorithms that usually maintain two-stage learning strategies, i.e., learning a dictionary in an unsupervised way then followed by training a classifier. However, the treating dictionary learning and classifier training as separate stages may not produce both descriptive and discriminative models for objects. By contrast, our method is capable of constructing a dictionary that not only fully reflects the intrinsic manifold structure of the data, but also possesses discriminative power. This paper presents an optimization method to obtain such an optimal dictionary, associated sparse coding, and a classifier in an iterative process. Our experiments on a benchmark show our tracker attains outstanding performance compared with the state-of-the-art algorithms.

****

Self-Occlusions and Disocclusions in Causal Video Object Segmentation

Yanchao Yang, Ganesh Sundaramoorthi, Stefano Soatto; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4408-4416

We propose a method to detect disocclusion in video sequences of three-dimensional scenes and to partition the disoccluded regions into objects, defined by coherent deformation corresponding to surfaces in the scene. Our method infers deformation fields that are piecewise smooth by construction without the need for an explicit regularizer and the associated choice of weight. It then partitions the disoccluded region and groups its components with objects by leveraging on the complementarity of motion and appearance cues: Where appearance changes within an object, motion can usually be reliably inferred and used for grouping. Where appearance is close to constant, it can be used for grouping directly. We integrate both cues in an energy minimization framework, incorporate prior assumptions explicitly into the energy, and propose a numerical scheme.

****

Large Displacement 3D Scene Flow With Occlusion Reasoning

Andrei Zanfir, Cristian Sminchisescu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4417-4425

3D motion estimation is a fundamental problem with many computer vision applications. With the emergence of modern, affordable and increasingly accurate RGB-D sensors, single view approaches for estimating 3D motion, also known as scene flow, are becoming popular. In this paper we propose a novel coarse to fine correspondence-based scene flow approach to account for the effects of large displacements and to model occlusion, based on explicit geometric reasoning. Our methodology enforces piecewise motion rigidity at the level of the depth point cloud without explicitly smoothing the parameters of adjacent neighborhoods. By integrating all geometric and photometric components in a single, consistent, occlusion-aware energy model our method is able to deal with fast motions and large occlusions areas, as present in challenging datasets like MPI Sintel Flow Dataset, which have recently been augmented with depth information. By explicitly modeling large displacements and occlusion, we can now more successfully work with difficult sequences which cannot be currently processed by state of the art scene flow methods that rely on small inter-frame motion assumptions. We also show that by leveraging depth information, we can obtain superior correspondence fields compared to the best state of the art large-displacement (2D) optical flow methods.
****

Co-Interest Person Detection From Multiple Wearable Camera Videos
Yuewei Lin, Kareem Abdelfatah, Youjie Zhou, Xiaochuan Fan, Hongkai Yu, Hui Qian, Song Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4426-4434

Wearable cameras, such as Google Glass and Go Pro, enable video data collection over larger areas and from different views. In this paper, we tackle a new problem of locating the co-interest person (CIP), i.e., the one who draws attention from most camera wearers, from temporally synchronized videos taken by multiple wearable cameras. Our basic idea is to exploit the motion patterns of people and use them to correlate the persons across different videos, instead of performing appearance-based matching as in traditional video co-segmentation/localization. This way, we can identify CIP even if a group of people with similar appearance are present in the view. More specifically, we detect a set of persons on each frame as the candidates of the CIP and then build a Conditional Random Field (CRF) model to select the one with consistent motion patterns in different videos and high spacial-temporal consistency in each video. We collect three sets of wearable-camera videos for testing the proposed algorithm. All the involved people have similar appearances in the collected videos and the experiments demonstrate the effectiveness of the proposed algorithm.
****

Sparse Dynamic 3D Reconstruction From Unsynchronized Videos
Enliang Zheng, Dinghuang Ji, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4435-4443

We target the sparse 3D reconstruction of dynamic objects observed by multiple unsynchronized video cameras with unknown temporal overlap. To this end, we develop a framework to recover the unknown structure without sequencing information across video sequences. Our proposed compressed sensing framework poses the estimation of 3D structure as the problem of dictionary learning. Moreover, we define our dictionary as the temporally varying 3D structure, while we define local sequencing information in terms of the sparse coefficients describing a locally linear 3D structural interpolation. Our formulation optimizes a biconvex cost function that leverages a compressed sensing formulation and enforces both structural dependency coherence across video streams, as well as motion smoothness across estimates from common video sources. Experimental results demonstrate the effectiveness of our approach in both synthetic data and captured imagery.
****

Category-Blind Human Action Recognition: A Practical Recognition System
Wenbo Li, Longyin Wen, Mooi Choo Chuah, Siwei Lyu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4444-4452

Existing human action recognition systems for 3D sequences obtained from the depth camera are designed to cope with only one action category, either single-person action or two-person interaction, and are difficult to be extended to scenari os where both action categories co-exist. In this paper, we propose the category-blind human recognition method (CHARM) which can recognize a human action without making assumptions of the action category. In our CHARM approach, we represent a human action (either a single-person action or a two-person interaction) class using a co-occurrence of motion primitives. Subsequently, we classify an action instance based on matching its motion primitive co-occurrence patterns to each class representation. The matching task is formulated as maximum clique problems. We conduct extensive evaluations of CHARM using three datasets for single-person actions, two-person interactions, and their mixtures. Experimental results show that CHARM performs favorably when compared with several state-of-the-art single-person action and two-person interaction based methods without making explicit assumptions of action category.

****

Temporal Subspace Clustering for Human Motion Segmentation

Sheng Li, Kang Li, Yun Fu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4453-4461

Subspace clustering is an effective technique for segmenting data drawn from multiple subspaces. However, for time series data (e.g., human motion), exploiting temporal information is still a challenging problem. We propose a novel temporal subspace clustering (TSC) approach in this paper. We improve the subspace clustering technique from two aspects. First, a temporal Laplacian regularization is designed, which encodes the sequential relationships in time series data. Second, to obtain expressive codings, we learn a non-negative dictionary from data. An efficient optimization algorithm is presented to jointly learn the representation codings and dictionary. After constructing an affinity graph using the codings, multiple temporal segments can be grouped via spectral clustering. Experimental results on three action and gesture datasets demonstrate the effectiveness of our approach. In particular, TSC significantly improves the clustering accuracy, compared to the state-of-the-art subspace clustering methods.

****

Weakly-Supervised Alignment of Video With Text

Piotr Bojanowski, Remi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4462-4470

Suppose that we are given a set of videos, along with natural language descriptions in the form of multiple sentences (e.g., manual annotations, movie scripts, sport summaries etc.), and that these sentences appear in the same temporal order as their visual counterparts. We propose in this paper a method for aligning the two modalities, i.e., automatically providing a time (frame) stamp for every sentence. Given vectorial features for both video and text, this can be cast as a temporal assignment problem, with an implicit linear mapping between the two feature modalities. We formulate this problem as an integer quadratic program, and solve its continuous convex relaxation using an efficient conditional gradient algorithm. Several rounding procedures are proposed to construct the final integer solution. After demonstrating significant improvements over the state of the art on the related task of aligning video with symbolic labels, we evaluate our method on a challenging dataset of videos with associated textual descriptions, and explore bag-of-words and continuous representations for text.

****

Learning Temporal Embeddings for Complex Video Analysis

Vignesh Ramanathan, Kevin Tang, Greg Mori, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4471-4479

In this paper, we propose to learn temporal embeddings of video frames for complex video analysis. Large quantities of unlabeled video data can be easily obtained from the Internet. These videos possess the implicit weak label that they are sequences of temporally and semantically coherent images. We leverage this information to learn temporal embeddings for video frames by associating frames with the temporal context that they appear in. To do this, we propose a scheme for incorporating temporal context based on past and future frames in videos, and compare this to other contextual representations. In addition, we show how data aug mentation using multi-resolution samples and hard negatives helps to significantly improve the quality of the learned embeddings. We evaluate various design decisions for learning temporal embeddings, and show that our embeddings can improve performance for multiple video tasks such as retrieval, classification, and temporal order recovery in unconstrained Internet video.

****

Unsupervised Semantic Parsing of Video Collections

Ozan Sener, Amir R. Zamir, Silvio Savarese, Ashutosh Saxena; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4480-4488

Human communication typically has an underlying structure. This is reflected in the fact that in many user generated videos, a starting point, ending, and certain objective steps between these two can be identified. In this paper, we propose a method for parsing a video into such semantic steps in an unsupervised way. The proposed method is capable of providing a semantic ``storyline'' of the video composed of its objective steps. We accomplish this utilizing both visual and language cues in a joint generative model. The proposed method can also provide a textual description for each of identified semantic steps and video segments. We evaluate this method on a large number of complex YouTube videos and show results of unprecedented quality for this new and impactful problem.

****

Learning Spatiotemporal Features With 3D Convolutional Networks

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497

We propose a simple, yet effective approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large scale supervised video dataset. Our findings are three-fold: 1) 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets; 2) A homogeneous architecture with small 3x3x3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets; and 3) Our learned features, namely C3D (Convolutional 3D), with a simple linear classifier outperform state-of-the-art methods on 4 different benchmarks and are comparable with current best methods on the other 2 benchmarks. In addition, the features are compact: achieving 52.8% accuracy on UCF101 dataset with only 10 dimensions and also very efficient to compute due to the fast inference of ConvNets. Finally, they are conceptually very simple and easy to train and use.

****

Temporal Perception and Prediction in Ego-Centric Video

Yipin Zhou, Tamara L. Berg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4498-4506

Given a video of an activity, can we predict what will happen next? In this paper we explore two simple tasks related to temporal prediction in egocentric videos of everyday activities. We provide both human experiments to understand how well people can perform on these tasks and computational models for prediction. Experiments indicate that humans and computers can do well on temporal prediction and that personalization to a particular individual or environment provides significantly increased performance. Developing methods for temporal prediction could have far reaching benefits for robots or intelligent agents to anticipate what a person will do, before they do it.

****

Describing Videos by Exploiting Temporal Structure

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4507-4515

Recent progress in using recurrent neural networks (RNNs) for image description has motivated the exploration of their application for video description. However, while images are static, working with videos requires modeling their dynamic temporal structure and then properly integrating that information into a natural language description model. In this context, we propose an approach that successfully takes into account both the local and global temporal structure of videos to produce descriptions. First, our approach incorporates a spatial temporal 3-D convolutional neural network (3-D CNN) representation of the short temporal dynamics. The 3-D CNN representation is trained on video action recognition tasks, so as to produce a representation that is tuned to human motion and behavior. Second we propose a temporal attention mechanism that allows to go beyond local temporal modeling and learns to automatically select the most relevant temporal segments given the text-generating RNN. Our approach exceeds the current state-of-art for both BLEU and METEOR metrics on the Youtube2Text dataset. We also present results on a new, larger and more challenging dataset of paired video and natural language descriptions.

****

Person Re-Identification With Discriminatively Trained Viewpoint Invariant Dictionaries

Srikrishna Karanam, Yang Li, Richard J. Radke; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4516-4524

This paper introduces a new approach to address the person re-identification problem in cameras with non-overlapping fields of view. Unlike previous approaches that learn Mahalanobis-like distance metrics in some transformed feature space, we propose to learn a dictionary that is capable of discriminatively and sparsely encoding features representing different people. Our approach directly addresses two key challenges in person re-identification: viewpoint variations and discriminability. First, to tackle viewpoint and associated appearance changes, we learn a single dictionary to represent both gallery and probe images in the training phase. We then discriminatively train the dictionary by enforcing explicit constraints on the associated sparse representations of the feature vectors. In the testing phase, we re-identify a probe image by simply determining the gallery image that has the closest sparse representation to that of the probe image in the Euclidean sense. Extensive performance evaluations on three publicly available multi-shot re-identification datasets demonstrate the advantages of our algorithm over several state-of-the-art dictionary learning, temporal sequence matching, and spatial appearance and metric learning based techniques.

****

Storyline Representation of Egocentric Videos With an Applications to Story-Based Search

Bo Xiong, Gunhee Kim, Leonid Sigal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4525-4533

Egocentric videos are a valuable source of information as a daily log of our lives. However, large fraction of egocentric video content is typically irrelevant and boring to re-watch. It is an agonizing task, for example, to manually search for the moment when your daughter first met Mickey Mouse from hours-long egocentric videos taken at Disneyland. Although many summarization methods have been successfully proposed to create concise representations of videos, in practice, the value of the subshots to users may change according to their immediate preference/mood; thus summaries with fixed criteria may not fully satisfy users' various search intents. To address this, we propose a storyline representation that expresses an egocentric video as a set of jointly inferred, through MRF inference, story elements comprising of actors, locations, supporting objects and events, depicted on a timeline. We construct such a storyline with very limited annotation data (a list of map locations and weak knowledge of what events may be possible at each location), by bootstrapping the process with data obtained through focused Web image and video searches. Our representation promotes story-based search with queries in the form of AND-OR graphs, which span any subset of story elements and their spatio-temporal composition. We show effectiveness of our approach on a set of unconstrained YouTube egocentric videos of visits to Disneyland.

****

Sequence to Sequence - Video to Text

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4534-4542

Real-world videos often have complex dynamics; methods for generating open-domai n video descriptions should be senstive to temporal structure and allow both input (sequence of frames) and output (sequence of words) of variable length. To approach this problem we propose a novel end-to-end sequence-to-sequence model to generate captions for videos. For this we exploit recurrent neural networks, specifically LSTMs, which have demonstrated state-of-the-art performance in image caption generation. Our LSTM model is trained on video-sentence pairs and learns to associate a sequence of video frames to a sequence of words in order to generate a description of the event in the video clip. Our model naturally is able to learn the temporal structure of the sequence of frames as well as the sequence model of the generated sentences, i.e. a language model. We evaluate several variants of our model that exploit different visual features on a standard set of YouTube videos and two movie description datasets (M-VAD and MPII-MD).

****

Context Aware Active Learning of Activity Recognition Models

Mahmudul Hasan, Amit K. Roy-Chowdhury; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4543-4551

Activity recognition in video has recently benefited from the use of the context e.g., inter-relationships among the activities and objects. However, these approaches require data to be labeled and entirely available at the outset. In contrast, we formulate a continuous learning framework for context aware activity recognition from unlabeled video data which has two distinct advantages over most existing methods. First, we propose a novel active learning technique which not only exploits the informativeness of the individual activity instances but also utilizes their contextual information during the query selection process; this leads to significant reduction in expensive manual annotation effort. Second, the learned models can be adapted online as more data is available. We formulate a conditional random field (CRF) model that encodes the context and devise an information theoretic approach that utilizes entropy and mutual information of the nodes to compute the set of most informative query instances, which need to be labeled by a human. These labels are combined with graphical inference techniques for incrementally updating the model as new videos come in. Experiments on four challenging datasets demonstrate that our framework achieves superior performance with significantly less amount of manual labeling.

****

Action Recognition by Hierarchical Mid-Level Action Elements

Tian Lan, Yuke Zhu, Amir Roshan Zamir, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4552-4560

Realistic videos of human actions exhibit rich spatiotemporal structures at multiple levels of granularity: an action can always be decomposed into multiple finer-grained elements in both space and time. To capture this intuition, we propose to represent videos by a hierarchy of mid-level action elements (MAEs), where each MAE corresponds to an action-related spatiotemporal segment in the video. We introduce an unsupervised method to generate this representation from videos. Our method is capable of distinguishing action-related segments from background segments and representing actions at multiple spatiotemporal resolutions. Given a set of spatiotemporal segments generated from the training data, we introduce a discriminative clustering algorithm that automatically discovers MAEs at multiple levels of granularity. We develop structured models that capture a rich set of spatial, temporal and hierarchical relations among the segments, where the action label and multiple levels of MAE labels are jointly inferred. The proposed model achieves state-of-the-art performance in multiple action recognition benchmarks. Moreover, we demonstrate the effectiveness of our model in real-world applications such as action recognition in large-scale untrimmed videos and action parsing.

****

Selecting Relevant Web Trained Concepts for Automated Event Retrieval

Bharat Singh, Xintong Han, Zhe Wu, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4561-4569

Complex event retrieval is a challenging research problem, especially when no tr aining videos are available. An alternative to collecting training videos is to train a large semantic concept bank a priori. Given a text description of an event, event retrieval is performed by selecting concepts linguistically related to the event description and fusing the concept responses on unseen videos. However, defining an exhaustive concept lexicon and pre-training it requires vast computational resources. Therefore, recent approaches automate concept discovery and training by leveraging large amounts of weakly annotated web data. Compact visually salient concepts are automatically obtained by the use of concept pairs or, more generally, n-grams. However, not all visually salient n-grams are necessarily useful for an event query--some combinations of concepts may be visually compact but irrelevant--and this drastically affects performance. We propose an event retrieval algorithm that constructs pairs of automatically discovered concepts and then prunes those concepts that are unlikely to be helpful for retrieval. Pruning depends both on the query and on the specific video instance being evaluated. Our approach also addresses calibration and domain adaptation issues that arise when applying concept detectors to unseen videos. We demonstrate large improvements over other vision based systems on the TRECVID MED 13 dataset.

****

Beyond Covariance: Feature Representation With Nonlinear Kernel Matrices

Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, Wanqing Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4570-4578

Covariance matrix has recently received increasing attention in computer vision by leveraging Riemannian geometry of symmetric positive-definite (SPD) matrices. Originally proposed as a region descriptor, it has now been used as a generic representation in various recognition tasks. However, covariance matrix has shortcomings such as being prone to be singular, limited capability in modeling complicated feature relationship, and having a fixed form of representation. This paper argues that more appropriate SPD-matrix-based representations shall be explored to achieve better recognition. It proposes an open framework to use the kernel matrix over feature dimensions as a generic representation and discusses its properties and advantages. The proposed framework significantly elevates covariance representation to the unlimited opportunities provided by this new representation. Experimental study shows that this representation consistently outperforms its covariance counterpart on various visual recognition tasks. In particular, it achieves significant improvement on skeleton-based human action recognition, demonstrating the state-of-the-art performance over both the covariance and the existing non-covariance representations.

****

Multiresolution Hierarchy Co-Clustering for Semantic Segmentation in Sequences With Small Variations

David Varas, Monica Alfaro, Ferran Marques; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4579-4587

This paper presents a co-clustering technique that, given a collection of images and their hierarchies, clusters nodes from these hierarchies to obtain a coherent multiresolution representation of the image collection. We formalize the co-clustering as Quadratic Semi-Assignment Problem and solve it with a linear programming relaxation approach that makes effective use of information from hierarchies. Initially, we address the problem of generating an optimal, coherent partition per image and, afterwards, we extend this method to a multiresolution framework. Finally, we particularize this framework to an iterative multiresolution video segmentation algorithm in sequences with small variations. We evaluate the algorithm on the Video Occlusion/Object Boundary Detection Dataset, showing that it produces state-of-the-art results in these scenarios.

****

Objects2action: Classifying and Localizing Actions Without Any Video Example

Mihir Jain, Jan C. van Gemert, Thomas Mensink, Cees G. M. Snoek; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4588-4596

The goal of this paper is to recognize actions in video without the need for examples. Different from traditional zero-shot approaches we do not demand the design and specification of attribute classifiers and class-to-attribute mappings to allow for transfer from seen classes to unseen classes. Our key contribution is objects2action, a semantic word embedding that is spanned by a skip-gram model of thousands of object categories. Action labels are assigned to an object encoding of unseen video based on a convex combination of action and object affinities. Our semantic embedding has three main characteristics to accommodate for the specifics of actions. First, we propose a mechanism to exploit multiple-word descriptions of actions and objects. Second, we incorporate the automated selection of the most responsive objects per action. And finally, we demonstrate how to extend our zero-shot approach to the spatio-temporal localization of actions in video. Experiments on four action datasets demonstrate the potential of our approach.

****

Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks

Lin Sun, Kui Jia, Dit-Yan Yeung, Bertram E. Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4597-4605

Human actions in video sequences are three-dimensional (3D) spatio-temporal signals characterizing both the visual appearance and motion dynamics of the involved humans and objects. Inspired by the success of convolutional neural networks (CNN) for image classification, recent attempts have been made to learn 3D CNNs for recognizing human actions in videos. However, partly due to the high complexity of training 3D convolution kernels and the need for large quantities of training videos, only limited success has been reported. This has triggered us to investigate in this paper a new deep architecture which can handle 3D signals more effectively. Specifically, we propose factorized spatio-temporal convolutional networks (FstCN) that factorize the original 3D convolution kernel learning as a sequential process of learning 2D spatial kernels in the lower layers (called spatial convolutional layers), followed by learning 1D temporal kernels in the upper layers (called temporal convolutional layers). We introduce a novel transformation and permutation operator to make factorization in FstCN possible. Moreover, to address the issue of sequence alignment, we propose an effective training and inference strategy based on sampling multiple video clips from a given action video sequence. We have tested FstCN on two commonly used benchmark datasets (UCF-101 and HMDB-51). Without using auxiliary training videos to boost the performance, FstCN outperforms existing CNN based methods and achieves comparable performance with a recent method that benefits from using auxiliary training videos.

****

Bayesian Non-Parametric Inference for Manifold Based MoCap Representation

Fabrizio Natola, Valsamis Ntouskos, Marta Sanzari, Fiora Pirri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4606-4614

We propose a novel approach to human action recognition, with motion capture data (MoCap), based on grouping sub-body parts. By representing configurations of actions as manifolds, joint positions are mapped on a subspace via principal geodesic analysis. The reduced space is still highly informative and allows for classification based on a non-parametric Bayesian approach, generating behaviors for each sub-body part. Having partitioned the set of joints, poses relative to a sub-body part are exchangeable, given a specified prior and can elicit, in principle, infinite behaviors. The generation of these behaviors is specified by a Dirichlet process mixture. We show with several experiments that the recognition gives very promising results, outperforming methods requiring temporal alignment.

****

Semantic Video Entity Linking Based on Visual Content and Metadata

Yuncheng Li, Xitong Yang, Jiebo Luo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4615-4623

Video entity linking, which connects online videos to the related entities in a semantic knowledge base, can enable a wide variety of video based applications including video retrieval and video recommendation. Most existing systems for video entity linking rely on video metadata. In this paper, we propose to exploit video visual content to improve video entity linking. In the proposed framework, videos are first linked to entity candidates using a text-based method. Next, th

e entity candidates are verified and reranked according to visual content. In order to properly handle large variations in visual content matching, we propose to use Multiple Instance Metric Learning to learn a "set to sequence'' metric for this specific matching problem. To evaluate the proposed framework, we collect and annotate 1912 videos crawled from the YouTube open API. Experiment results have shown consistent gains by the proposed framework over several strong baselines.
****
Love Thy Neighbors: Image Annotation by Exploiting Image Metadata
Justin Johnson, Lamberto Ballan, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4624-4632
Some images that are difficult to recognize on their own may become more clear in the context of a neighborhood of related images with similar social-network metadata. We build on this intuition to improve multilabel image annotation. Our model uses image metadata nonparametrically to generate neighborhoods of related images using Jaccard similarities, then uses a deep neural network to blend visual information from the image and its neighbors. Prior work typically models image metadata parametrically; in contrast, our nonparametric treatment allows our model to perform well even when the vocabulary of metadata changes between training and testing. We perform comprehensive experiments on the NUS-WIDE dataset, where we show that our model outperforms state-of-the-art methods for multilabel image annotation even when our model is forced to generalize to new types of metadata.
****
Unsupervised Extraction of Video Highlights Via Robust Recurrent Auto-Encoders
Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, Baining Guo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4633-4641
With the growing popularity of short-form video sharing platforms such as Instagram and Vine, there has been an increasing need for techniques that automatically extract highlights from video. Whereas prior works have approached this problem with heuristic rules or supervised learning, we present an unsupervised learning approach that takes advantage of the abundance of user-edited videos on social media websites such as YouTube. Based on the idea that the most significant sub-events within a video class are commonly present among edited videos while less interesting ones appear less frequently, we identify the significant sub-events via a robust recurrent auto-encoder trained on a collection of user-edited videos queried for each particular class of interest. The auto-encoder is trained using a proposed shrinking exponential loss function that makes it robust to noise in the web-crawled training data, and is configured with bidirectional long short term memory (LSTM) cells to better model the temporal structure of highlight segments. Different from supervised techniques, our method can infer highlights using only a set of downloaded edited videos, without also needing their pre-edited counterparts which are rarely available online. Extensive experiments indicate the promise of our proposed solution in this challenging unsupervised setting.
****
Learning Visual Clothing Style With Heterogeneous Dyadic Co-Occurrences
Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, Serge Belongie; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4642-4650
With the rapid proliferation of smart mobile devices, users now take millions of photos every day. These include large numbers of clothing and accessory images. We would like to answer questions like `What outfit goes well with this pair of shoes?' To answer these types of questions, one has to go beyond learning visual similarity and learn a visual notion of compatibility across categories. In this paper, we propose a novel learning framework to help answer these types of questions. The main idea of this framework is to learn a feature transformation from images of items into a latent space that expresses compatibility. For the feature transformation, we use a Siamese Convolutional Neural Network (CNN) archite

cture, where training examples are pairs of items that are either compatible or incompatible. We model compatibility based on co-occurrence in large-scale user behavior data; in particular co-purchase data from Amazon.com. To learn cross-category fit, we introduce a strategic method to sample training data, where pairs of items are heterogeneous dyads, i.e., the two elements of a pair belong to different high-level categories. While this approach is applicable to a wide variety of settings, we focus on the representative problem of learning compatible clothing style. Our results indicate that the proposed framework is capable of learning semantic information about visual style and is able to generate outfits of clothes, with items from different categories, that go well together.

****

Text Flow: A Unified Text Detection System in Natural Scene Images

Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, Chew Lim Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4651-4659

The prevalent scene text detection approach follows four sequential steps comprising character candidate detection, false character candidate removal, text line extraction, and text line verification. However, errors occur and accumulate throughout each of these sequential steps which often lead to low detection performance. To address these issues, we propose a unified scene text detection system, namely Text Flow, by utilizing the minimum cost (min-cost) flow network model. With character candidates detected by cascade boosting, the min-cost flow network model integrates the last three sequential steps into a single process which solves the error accumulation problem at both character level and text line level effectively. The proposed technique has been tested on three public datasets, i.e, ICDAR2011 dataset, ICDAR2013 dataset and a multilingual dataset and it outperforms the state-of-the-art methods on all three datasets with much higher recall and F-score. The good performance on the multilingual dataset shows that the proposed technique can be used for the detection of texts in different languages.

****

Uncovering Interactions and Interactors: Joint Estimation of Head, Body Orientation and F-Formations From Surveillance Videos

Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulo, Narendra Ahuja, Oswald Lanz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4660-4668

We present a novel approach for jointly estimating tar- gets' head, body orientations and conversational groups called F-formations from a distant social scene (e.g., a cocktail party captured by surveillance cameras). Differing from related works that have (i) coupled head and body pose learning by exploiting the limited range of orientations that the two can jointly take, or (ii) determined F-formations based on the mutual head (but not body) orientations of in- teractors, we present a unified framework to jointly infer both (i) and (ii). Apart from exploiting spatial and orien- tation relationships, we also integrate cues pertaining to temporal consistency and occlusions, which are beneficial while handling low-resolution data under surveillance set- tings. Efficacy of the joint inference framework reflects via increased head, body pose and F-formation estimation ac- curacy over the state-of-the-art, as confirmed by extensive experiments on two social datasets.

****

Generating Notifications for Missing Actions: Don't Forget to Turn the Lights Off!

Bilge Soran, Ali Farhadi, Linda Shapiro; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4669-4677

We all have experienced forgetting habitual actions among our daily activities. For example, we probably have forgotten to turn the lights off before leaving a room or turn the stove off after cooking. In this paper, we propose a solution to the problem of issuing notifications on actions that may be missed. This involves learning about interdependencies between actions and being able to predict an ongoing action while segmenting the input video stream. In order to show a pro of of concept, we collected a new egocentric dataset, in which people wear a camera while making lattes. We show promising results on the extremely challenging task of issuing correct and timely reminders. We also show that our model reliably segments the actions, while predicting the ongoing one when only a few frames from the beginning of the action are observed. The overall prediction accuracy is 46.2% when only 10 frames of an action are seen (2/3 of a sec). Moreover, the overall recognition and segmentation accuracy is shown to be 72.7% when the whole activity sequence is observed. Finally, the online prediction and segmentation accuracy is 68.3% when the prediction is made at every time step.

****

Partial Person Re-Identification

Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, Shaogang Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4678-4686

We address a new partial person re-identification (re-id) problem, where only a partial observation of a person is available for matching across different non-overlapping camera views. This differs significantly from the conventional person re-id setting where it is assumed that the full body of a person is detected and aligned. To solve this more challenging and realistic re-id problem without the implicit assumption of manual body-parts alignment, we propose a matching framework consisting of 1) a local patch-level matching model based on a novel sparse representation classification formulation with explicit patch ambiguity modelling, and 2) a global part-based matching model providing complementary spatial layout information. Our framework is evaluated on a new partial person re-id dataset as well as two existing datasets modified to include partial person images. The results show that the proposed method outperforms significantly existing re-id methods as well as other partial visual matching methods.

****

Shape Interaction Matrix Revisited and Robustified: Efficient Subspace Clustering With Corrupted and Incomplete Data

Pan Ji, Mathieu Salzmann, Hongdong Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4687-4695

The Shape Interaction Matrix (SIM) is one of the earliest approaches to performing subspace clustering (i.e., separating points drawn from a union of subspaces). In this paper, we revisit the SIM and reveal its connections to several recent subspace clustering methods. Our analysis lets us derive a simple, yet effective algorithm to robustify the SIM and make it applicable to realistic scenarios where the data is corrupted by noise. We justify our method by intuitive examples and the matrix perturbation theory. We then show how this approach can be extended to handle missing data, thus yielding an efficient and general subspace clustering algorithm. We demonstrate the benefits of our approach over state-of-the-art subspace clustering methods on several challenging motion segmentation and face clustering problems, where the data includes corruptions and missing measurements.

****

Multiple Hypothesis Tracking Revisited

Chanho Kim, Fuxin Li, Arridhana Ciptadi, James M. Rehg; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4696-4704

This paper revisits the classical multiple hypotheses tracking (MHT) algorithm in a tracking-by-detection framework. The success of MHT largely depends on the ability to maintain a small list of potential hypotheses, which can be facilitated with the accurate object detectors that are currently available. We demonstrate that a classical MHT implementation from the 90's can come surprisingly close to the performance of state-of-the-art methods on standard benchmark datasets. In order to further utilize the strength of MHT in exploiting higher-order information, we introduce a method for training online appearance models for each track hypothesis. We show that appearance models can be learned efficiently via a regularized least squares framework, requiring only a few extra operations for each hypothesis branch. We obtain state-of-the-art results on popular tracking-by-detection datasets such as PETS and the recent MOT challenge.

****

Learning to Track: Online Multi-Object Tracking by Decision Making

Yu Xiang, Alexandre Alahi, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4705-4713

Online Multi-Object Tracking (MOT) has wide applications in time-critical video analysis scenarios, such as robot navigation and autonomous driving. In tracking-by-detection, a major challenge of online MOT is how to robustly associate noisy object detections on a new video frame with previously tracked objects. In this work, we formulate the online MOT problem as decision making in Markov Decision Processes (MDPs), where the lifetime of an object is modeled with a MDP. Learning a similarity function for data association is equivalent to learning a policy for the MDP, and the policy learning is approached in a reinforcement learning fashion which benefits from both advantages of offline-learning and online-learning for data association. Moreover, our framework can naturally handle the birth/death and appearance/disappearance of targets by treating them as state transitions in the MDP while leveraging existing online single object tracking methods. We conduct experiments on the MOT Benchmark to verify the effectiveness of our method.

****