Globally-Optimal Inlier Set Maximisation for Simultaneous Camera Pose and Feature Correspondence

Dylan Campbell, Lars Petersson, Laurent Kneip, Hongdong Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1-10

Estimating the 6-DoF pose of a camera from a single image relative to a pre-computed 3D point-set is an important task for many computer vision applications. Perspective-n-Point (PnP) solvers are routinely used for camera pose estimation, provided that a good quality set of 2D-3D feature correspondences are known beforehand. However, finding optimal correspondences between 2D key-points and a 3D point-set is non-trivial, especially when only geometric (position) information is known. Existing approaches to the simultaneous pose and correspondence problem use local optimisation, and are therefore unlikely to find the optimal solution without a good pose initialisation, or introduce restrictive assumptions. Since a large proportion of outliers are common for this problem, we instead propose a globally-optimal inlier set cardinality maximisation approach which jointly estimates optimal camera pose and optimal correspondences. Our approach employs branch-and-bound to search the 6D space of camera poses, guaranteeing global optimality without requiring a pose prior. The geometry of SE(3) is used to find novel upper and lower bounds for the number of inliers and local optimisation is integrated to accelerate convergence. The evaluation empirically supports the optimality proof and shows that the method performs much more robustly than existing approaches, including on a large-scale outdoor data-set.

****

Robust Pseudo Random Fields for Light-Field Stereo Matching

Chao-Tsung Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 11-19

Markov Random Fields are widely used to model light-field stereo matching problems. However, most previous approaches used fixed parameters and did not adapt to light-field statistics. Instead, they explored explicit vision cues to provide local adaptability and thus enhanced depth quality. But such additional assumptions could end up confining their applicability, e.g. algorithms designed for dense light fields are not suitable for sparse ones. In this paper, we develop an empirical Bayesian framework--Robust Pseudo Random Field--to explore intrinsic statistical cues for broad applicability. Based on pseudo-likelihood, it applies soft expectation-maximization (EM) for good model fitting and hard EM for robust depth estimation. We introduce novel pixel difference models to enable such adaptability and robustness simultaneously. We also devise an algorithm to employ this framework on dense, sparse, and even denoised light fields. Experimental results show that it estimates scene-dependent parameters robustly and converges quickly. In terms of depth accuracy and computation speed, it also outperforms state-of-the-art algorithms constantly.

****

A Lightweight Approach for On-The-Fly Reflectance Estimation

Kihwan Kim, Jinwei Gu, Stephen Tyree, Pavlo Molchanov, Matthias Niessner, Jan Kautz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 20-28

Estimating surface reflectance (BRDF) is one key component for complete 3D scene capture, with wide applications in virtual reality, augmented reality, and human computer interaction. Prior work is either limited to controlled environments (e.g., gonioreflectometers, light stages or multi-camera domes), or requires the joint optimization of shape, illumination, and reflectance, which is often computationally too expensive (e.g., hours of running time) for real-time applications. Moreover, most prior work requires HDR images as input which further complicates the capture process. In this paper, we propose a lightweight, practical approach for surface reflectance estimation directly from 8-bit RGB images in real-time, which can be easily plugged into any 3D scanning-and-fusion system with a commodity RGBD sensor. Our method is learning-based, with an inference time of less than 90ms per scene and a model size of less than 340K bytes. We propose two novel network architectures, HemiCNN and Grouplet, to deal with the unstructured input data from multiple viewpoints under unknown illumination. We further des ign a loss function to resolve the color-constancy and scale ambiguity. In addition, we have created a large synthetic dataset, SynBRDF, which comprises a total of 500K RGBD images rendered with a physically-based ray tracer under a variety of natural illumination, covering 5000 materials and 5000 shapes. SynBRDF is the first large-scale benchmark dataset for reflectance estimation. Experiments on both synthetic data and real data show that the proposed method effectively recovers surface reflectance, and outperforms prior work for reflectance estimation in uncontrolled environments.

****

Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus

Runze Zhang, Siyu Zhu, Tian Fang, Long Quan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 29-38

The increasing scale of Structure-from-Motion is fundamentally limited by the conventional optimization framework for the all-in-one global bundle adjustment. In this paper, we propose a distributed approach to coping with this global bundle adjustment for very large scale Structure-from-Motion computation. First, we derive the distributed formulation from the classical optimization algorithm ADMM, Alternating Direction Method of Multipliers, based on the global camera consensus. Then, we analyze the conditions under which the convergence of this distributed optimization would be guaranteed. In particular, we adopt over-relaxation and self-adaption schemes to improve the convergence rate. After that, we propose to split the large scale camera-point visibility graph in order to reduce the communication overheads of the distributed computing. The experiments on both public large scale SfM data-sets and our very large scale aerial photo sets demonstrate that the proposed distributed method clearly outperforms the state-of-the-art method in efficiency and accuracy.

****

Practical Projective Structure From Motion (P2SfM)

Ludovic Magerand, Alessio Del Bue; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 39-47

This paper presents a solution to the Projective Structure from Motion (PSfM) problem able to deal efficiently with missing data, outliers and, for the first time, large scale 3D reconstruction scenarios. By embedding the projective depths into the projective parameters of the points and views, we decrease the number of unknowns to estimate and improve computational speed by optimizing standard linear Least Squares systems instead of homogeneous ones. In order to do so, we show that an extension of the linear constraints from the Generalized Projective Reconstruction Theorem can be transferred to the projective parameters, ensuring also a valid projective reconstruction in the process. We use an incremental approach that, starting from a solvable sub-problem, incrementally adds views and points until completion with a robust, outliers free, procedure. Experiments with simulated data shows that our approach is performing well, both in term of the quality of the reconstruction and the capacity to handle missing data and outliers with a reduced computational time. Finally, results on real datasets shows the ability of the method to be used in medium and large scale 3D reconstruction scenarios with high ratios of missing data (up to 98%).

****

Anticipating Daily Intention Using On-Wrist Motion Triggered Sensing

Tz-Ying Wu, Ting-An Chien, Cheng-Sheng Chan, Chan-Wei Hu, Min Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 48-56

Anticipating human intention by observing one's actions has many applications. For instance, picking up a cellphone, then a charger (actions) implies that one wants to charge the cellphone (intention). By anticipating the intention, an intelligent system can guide the user to the closest power outlet. We propose an on-wrist motion triggered sensing system for anticipating daily intentions, where the on-wrist sensors help us to persistently observe one's actions. The core of the system is a novel Recurrent Neural Network (RNN) and Policy Network (PN), where the RNN encodes visual and motion observation to anticipate intention, and the PN parsimoniously triggers the process of visual observation to reduce computation requirement. We jointly trained the whole network using policy gradient and cross-entropy loss. To evaluate, we collect the first daily "intention" dataset consisting of 2379 videos with 34 intentions and 164 unique action sequences. Our method achieves 92.68%, 90.85%, 97.56% accuracy on three users while processing only 29% of the visual observation on average.

****

Rethinking Reprojection: Closing the Loop for Pose-Aware Shape Reconstruction From a Single Image

Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 57-65

An emerging problem in computer vision is the reconstruction of 3D shape and pose of an object from a single image. Hitherto, the problem has been addressed through the application of canonical deep learning methods to regress from the image directly to the 3D shape and pose labels. These approaches, however, are problematic from two perspectives. First, they are minimizing the error between 3D shapes and pose labels - with little thought about the nature of this "label error" when reprojecting the shape back onto the image. Second, they rely on the onerous and ill-posed task of hand labeling natural images with respect to 3D shape and pose. In this paper we define the new task of pose-aware shape reconstruction from a single image, and we advocate that cheaper 2D annotations of objects silhouettes in natural images can be utilized. We design architectures of pose-aware shape reconstruction which reproject the predicted shape back on to the image using the predicted pose. Our evaluation on several object categories demonstrates the superiority of our method for predicting pose-aware 3D shapes from natural images.

****

End-To-End Learning of Geometry and Context for Deep Stereo Regression

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, Adam Bry; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 66-75

We propose a novel deep learning architecture for regressing disparity from a rectified pair of stereo images. We leverage knowledge of the problem's geometry to form a cost volume using deep feature representations. We learn to incorporate contextual information using 3-D convolutions over this volume. Disparity values are regressed from the cost volume using a proposed differentiable soft argmin operation, which allows us to train our method end-to-end to sub-pixel accuracy without any additional post-processing or regularization. We evaluate our method on the Scene Flow and KITTI datasets and on KITTI we set a new state-of-the-art benchmark, while being significantly faster than competing approaches.

****

Using Sparse Elimination for Solving Minimal Problems in Computer Vision

Janne Heikkila; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 76-84

Finding a closed form solution to a system of polynomial equations is a common problem in computer vision as well as in many other areas of engineering and science. Groebner basis techniques are often employed to provide the solution, but implementing an efficient Groebner basis solver to a given problem requires strong expertise in algebraic geometry. One can also convert the equations to a polynomial eigenvalue problem (PEP) and solve it using linear algebra, which is a more accessible approach for those who are not so familiar with algebraic geometry. In previous works PEP has been successfully applied for solving some relative pose problems in computer vision, but its wider exploitation is limited by the problem of finding a compact monomial basis. In this paper, we propose a new algorithm for selecting the basis that is in general more compact than the basis obtained with a state-of-the-art algorithm making PEP a more viable option for solving polynomial equations. Another contribution is that we present two minimal problems for camera self-calibration based on homography, and demonstrate experimentally using synthetic and real data that our algorithm can provide a numerically stable solution to the camera focal length from two homographies of unknown planar scene.

****

High-Resolution Shape Completion Using Deep Neural Networks for Global Structure and Local Geometry Inference

Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, Yizhou Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 85-93

We propose a data-driven method for recovering missing parts of 3D shapes. Our method is based on a new deep learning architecture consisting of two sub-networks: a global structure inference network and a local geometry refinement network. The global structure inference network incorporates a long short-term memorized context fusion module (LSTM-CF) that infers the global structure of the shape based on multi-view depth information provided as part of the input. It also includes a 3D fully convolutional (3DFCN) module that further enriches the global structure representation according to volumetric information in the input. Under the guidance of the global structure network, the local geometry refinement network takes as input local 3D patches around missing regions, and progressively produces a high-resolution, complete surface through a volumetric encoder-decoder architecture. Our method jointly trains the global structure inference and local geometry refinement networks in an end-to-end manner. We perform qualitative and quantitative evaluations on six object categories, demonstrating that our method outperforms existing state-of-the-art work on shape completion.

****

Temporal Tessellation: A Unified Approach for Video Analysis

Dotan Kaufman, Gil Levi, Tal Hassner, Lior Wolf; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 94-104

We present a general approach to video understanding, inspired by semantic transfer techniques that have been successfully used for 2D image analysis. Our method considers a video to be a 1D sequence of clips, each one associated with its own semantics. The nature of these semantics -- natural language captions or other labels -- depends on the task at hand. A test video is processed by forming correspondences between its clips and the clips of reference videos with known semantics, following which, reference semantics can be transferred to the test video. We describe two matching methods, both designed to ensure that (a) reference clips appear similar to test clips and (b), taken together, the semantics of the selected reference clips is consistent and maintains temporal coherence. We use our method for video captioning on the LSMDC'16 benchmark, video summarization on the SumMe and TVSum benchmarks, Temporal Action Detection on the Thumos2015 benchmark, and sound prediction on the Greatest Hits benchmark. Our method not only surpasses the state of the art, in four out of five benchmarks, but importantly, it is the only single method we know of that was successfully applied to such a diverse range of tasks.

****

Learning Policies for Adaptive Tracking With Deep Feature Cascades

Chen Huang, Simon Lucey, Deva Ramanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 105-114

Visual object tracking is a fundamental and time-critical vision task. Recent years have seen many shallow tracking methods based on real-time pixel-based correlation filters, as well as deep methods that have top performance but need a high-end GPU. In this paper, we learn to improve the speed of deep trackers without losing accuracy. Our fundamental insight is to take an adaptive approach, where easy frames are processed with cheap features (such as pixel values), while challenging frames are processed with invariant but expensive deep features. We formulate the adaptive tracking problem as a decision-making process, and learn an agent to decide whether to locate objects with high confidence on an early layer, or continue processing subsequent layers of a network. This significantly reduces the feed-forward cost for easy frames with distinct or slow-moving objects. We train the agent offline in a reinforcement learning fashion, and further demonstrate that learning all deep layers (so as to provide good features for adaptive tracking) can lead to near real-time average tracking speed of 23 fps on a single CPU while achieving state-of-the-art performance. Perhaps most tellingly, our approach provides a 100X speedup for almost 50% of the time, indicating the p ower of an adaptive approach.

****

Temporal Shape Super-Resolution by Intra-Frame Motion Encoding Using High-Fps Structured Light

Yuki Shiba, Satoshi Ono, Ryo Furukawa, Shinsaku Hiura, Hiroshi Kawasaki; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 115-123

One of the solutions of depth imaging of moving scene is to project a static pattern on the object and use just a single image for reconstruction. However, if the motion of the object is too fast with respect to the exposure time of the image sensor, patterns on the captured image are blurred and reconstruction fails. In this paper, we impose multiple projection patterns into each single captured image to realize temporal super resolution of the depth image sequences. With our method, multiple patterns are projected onto the object with higher fps than possible with a camera. In this case, the observed pattern varies depending on the depth and motion of the object, so we can extract temporal information of the scene from each single image. The decoding process is realized using a learning-based approach where no geometric calibration is needed. Experiments confirm the effectiveness of our method where sequential shapes are reconstructed from a single image. Both quantitative evaluations and comparisons with recent techniques were also conducted.

****

Real-Time Monocular Pose Estimation of 3D Objects Using Temporally Consistent Local Color Histograms

Henning Tjaden, Ulrich Schwanecke, Elmar Schomer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 124-132

We present a novel approach to 6DOF pose estimation and segmentation of rigid 3D objects using a single monocular RGB camera based on temporally consistent, local color histograms. We show that this approach outperforms previous methods in cases of cluttered backgrounds, heterogenous objects, and occlusions. The proposed histograms can be used as statistical object descriptors within a template matching strategy for pose recovery after temporary tracking loss e. g. caused by massive occlusion or if the object leaves the camera's field of view. The descriptors can be trained online within a couple of seconds moving a handheld object in front of a camera. During the training stage, our approach is already capable to recover from accidental tracking loss. We demonstrate the performance of our method in comparison to the state of the art in different challenging experiments including a popular public data set.

****

CAD Priors for Accurate and Flexible Instance Reconstruction

Tolga Birdal, Slobodan Ilic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 133-142

We present an efficient and automatic approach for accurate reconstruction of instances of big 3D objects from multiple, unorganized and unstructured point clouds, in presence of dynamic clutter and occlusions. In contrast to conventional scanning, where the background is assumed to be rather static, we aim at handling dynamic clutter where background drastically changes during the object scanning. Currently, it is tedious to solve this with available methods unless the object of interest is first segmented out from the rest of the scene. We address the problem by assuming the availability of a prior CAD model, roughly resembling the object to be reconstructed. This assumption almost always holds in applications such as industrial inspection or reverse engineering. With aid of this prior acting as a proxy, we propose a fully enhanced pipeline, capable of automatically detecting and segmenting the object of interest from scenes and creating a pose graph, online, with linear complexity. This allows initial scan alignment to the CAD model space, which is then refined without the CAD constraint to fully recover a high fidelity 3D reconstruction, accurate up to the sensor noise level. We also contribute a novel object detection method, local implicit shape models (LISM) and give a fast verification scheme. We evaluate our method on multiple datasets, demonstrating the ability to accurately reconstruct objects from small s izes up to 125m3.

****

Colored Point Cloud Registration Revisited

Jaesik Park, Qian-Yi Zhou, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 143-152

We present an algorithm for tightly aligning two colored point clouds. The key idea is to optimize a joint photometric and geometric objective that locks the alignment along both the normal direction and the tangent plane. We extend a photometric objective for aligning RGB-D images to point clouds, by locally parameterizing the point cloud with a virtual camera. Experiments demonstrate that our algorithm is more accurate and more robust than prior point cloud registration algorithms, including those that utilize color information. We use the presented algorithms to enhance a state-of-the-art scene reconstruction system. The accuracy of the resulting system is demonstrated on real-world scenes with accurate ground-truth models.

****

Learning Compact Geometric Features

Marc Khoury, Qian-Yi Zhou, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 153-161

We present an approach to learning features that represent the local geometry around a point in an unstructured point cloud. Such features play a central role in geometric registration, which supports diverse applications in robotics and 3D vision. Current state-of-the-art local features for unstructured point clouds have been manually crafted and none combines the desirable properties of precision, compactness, and robustness. We show that features with these properties can be learned from data, by optimizing deep networks that map high-dimensional histograms into low-dimensional Euclidean spaces. The presented approach yields a family of features, parameterized by dimension, that are both more compact and more accurate than existing descriptors.

****

Joint Layout Estimation and Global Multi-View Registration for Indoor Reconstruction

Jeong-Kyun Lee, Jaewon Yea, Min-Gyu Park, Kuk-Jin Yoon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 162-171

In this paper, we propose an approach to jointly solve scene layout estimation and global registration problems for accurate indoor 3D reconstruction. Given a sequence of range data, we build a set of scene fragments using KinectFusion and register them through pose graph optimization. Afterwards, we alternate layout estimation and layout-based global registration processes in iterative fashion to complement each other. We extract the scene layout through hierarchical agglomerative clustering and energy-based multi-model fitting in consideration of noisy measurements. Having the estimated scene layout in one hand, we register all the range data through the global iterative closest point algorithm where the positions of 3D points that belong to the layout such as walls and a ceiling are constrained to be close to the layout. We experimentally verify the proposed method with the publicly available synthetic and real-world datasets in both quantitative and qualitative ways.

****

A Geometric Framework for Statistical Analysis of Trajectories With Distinct Temporal Spans

Rudrasis Chakraborty, Vikas Singh, Nagesh Adluru, Baba C. Vemuri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 172-181

Analyzing data representing multifarious trajectories is central to the many fields in Science and Engineering; for example, trajectories representing a tennis serve, a gymnast's parallel bar routine, progression/remission of disease and so on. We present a novel geometric algorithm for performing statistical analysis of trajectories with distinct number of samples representing longitudinal (or temporal) data. A key feature of our proposal is that unlike existing schemes, our model is deployable in regimes where each participant provides a different number of acquisitions (trajectories have different number of sample points). To ach ieve this, we develop a novel method involving the parallel transport of the tangent vectors along each given trajectory to the starting point of the respective trajectories and then use the span of the matrix whose columns consist of these vectors, to construct a linear subspace in R^m. We then map these linear subspaces of R^m on to a single high dimensional hypersphere. This enables computing group statistics over trajectories by instead performing statistics on the hypersphere (equipped with a simpler geometry). Given a point on the hypersphere representing a trajectory, we also provide a "reverse mapping" algorithm to uniquely (under certain assumptions) reconstruct the subspace that corresponds to this point. Finally, by using existing algorithms for recursive Frechet mean and exact principal geodesic analysis on the hypersphere, we present several experiments on synthetic and real (vision and medical) data sets showing how group testing on such diversely sampled longitudinal data is possible by analyzing the reconstructed data in the subspace spanned by the first few PGs.

****

An Optimal Transportation Based Univariate Neuroimaging Index

Liang Mi, Wen Zhang, Junwei Zhang, Yonghui Fan, Dhruman Goradia, Kewei Chen, Eric M. Reiman, Xianfeng Gu, Yalin Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 182-191

The alterations of brain structures and functions have been considered closely correlated to the change of cognitive performance due to neurodegenerative diseases such as Alzheimer's disease. In this paper, we introduce a variational framework to compute the optimal transformation (OT) in 3D space and propose a univariate neuroimaging index based on OT to measure such alterations. We compute the OT from each image to a template and measure the Wasserstein distance between them. By comparing the distances from all the images to the common template, we obtain a concise and informative index for each image. Our framework makes use of the Newton's method, which reduces the computational cost and enables itself to be applicable to large-scale datasets. The proposed work is a generic approach and thus may be applicable to various volumetric brain images, including structural magnetic resonance (sMR) and fluorodeoxyglucose positron emission tomography (FDG-PET) images. In the classification between Alzheimer's disease patients and healthy controls, our method achieves an accuracy of 82.30% on the Alzheimer's Disease Neuroimaging Initiative (ADNI) baseline sMRI dataset and outperforms several other indices. On FDG-PET dataset, we boost the accuracy to 88.37% by leveraging pairwise Wasserstein distances. In a longitudinal study, we obtain a 5% significance with p-value = 0.0000113 in a t-test on FDG-PET. The results demonstrate a great potential of the proposed index for neuroimage analysis and the precision medicine research.

****

S3FD: Single Shot Scale-Invariant Face Detector

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan Z. Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 192-201

This paper presents a real-time face detector, named Single Shot Scale-invariant Face Detector (S3FD), which performs superiorly on various scales of faces with a single deep neural network, especially for small faces. Specifically, we try to solve the common problem that anchor-based detectors deteriorate dramatically as the objects become smaller. We make contributions in the following three aspects: 1) proposing a scale-equitable face detection framework to handle different scales of faces well. We tile anchors on a wide range of layers to ensure that all scales of faces have enough features for detection. Besides, we design anchor scales based on the effective receptive field and a proposed equal proportion interval principle; 2) improving the recall rate of small faces by a scale compensation anchor matching strategy; 3) reducing the false positive rate of small faces via a max-out background label. As a consequence, our method achieves state-of-the-art detection performance on all the common face detection benchmarks, including the AFW, PASCAL face, FDDB and WIDER FACE datasets, and can run at 36 FPS on a Nvidia Titan X (Pascal) for VGA-resolution images.

****

Amulet: Aggregating Multi-Level Convolutional Features for Salient Object Detection
Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, Xiang Ruan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 202-211
Fully convolutional neural networks (FCNs) have shown outstanding performance in many dense labeling problems. One key pillar of these successes is mining relevant information from features in convolutional layers. However, how to better aggregate multi-level convolutional feature maps for salient object detection is underexplored. In this work, we present Amulet, a generic aggregating multi-level convolutional feature framework for salient object detection. Our framework first integrates multi-level feature maps into multiple resolutions, which simultaneously incorporate coarse semantics and fine details. Then it adaptively learns to combine these feature maps at each resolution and predict saliency maps with the combined features. Finally, the predicted results are efficiently fused to generate the final saliency map. In addition, to achieve accurate boundary inference and semantic enhancement, edge-aware feature maps in low-level layers and the predicted results of low resolution features are recursively embedded into the learning framework. By aggregating multi-level convolutional features in this efficient and flexible manner, the proposed saliency model provides accurate salient object labeling. Comprehensive experiments demonstrate that our method performs favorably against state-of-the-art approaches in terms of near all compared evaluation metrics.
****
Learning Uncertain Convolutional Features for Accurate Saliency Detection
Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, Baocai Yin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 212-221
Deep convolutional neural networks (CNNs) have delivered superior performance in many computer vision tasks. In this paper, we propose a novel deep fully convolutional network model for accurate salient object detection. The key contribution of this work is to learn deep uncertain convolutional features (UCF), which encourage the robustness and accuracy of saliency detection. We achieve this via introducing a reformulated dropout (R-dropout) after specific convolutional layers to construct an uncertain ensemble of internal feature units. In addition, we propose an effective hybrid upsampling method to reduce the checkerboard artifacts of deconvolution operators in our decoder network. The proposed methods can also be applied to other deep convolutional networks. Compared with existing saliency detection methods, the proposed UCF model is able to incorporate uncertainties for more accurate object boundary inference. Extensive experiments demonstrate that our proposed saliency model performs favorably against state-of-the-art approaches. The uncertain feature learning mechanism as well as the upsampling method can significantly improve performance on other pixel-wise vision tasks.
****
Zero-Order Reverse Filtering
Xin Tao, Chao Zhou, Xiaoyong Shen, Jue Wang, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 222-230
In this paper, we study an unconventional but practically meaningful reversibility problem of commonly used image filters. We broadly define filters as operations to smooth images or to produce layers via global or local algorithms. And we raise the intriguingly problem if they are reservable to the status before filtering. To answer it, we present a novel strategy to understand general filter via contraction mappings on a metric space. A very simple yet effective zero-order algorithm is proposed. It is able to practically reverse most filters with low computational cost. We present quite a few experiments in the paper and supplementary file to thoroughly verify its performance. This method can also be generalized to solve other inverse problems and enables new applications.
****
Learning Blind Motion Deblurring
Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, Hendrik P. A. Lensch; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 231-240

As handheld video cameras are now commonplace and available in every smartphone images and videos can be recorded almost everywhere at any time. However, taking a quick shot frequently ends up in a blurry result due to unwanted camera shake during recording or moving objects in the scene. Removing these artifacts from the blurry recordings is a highly ill-posed problem as neither the sharp image nor the motion blur is known. Propagating information between multiple consecutive blurry observations can help to restore the desired sharp image or video. Solutions for blind deconvolution based on neural networks rely on a massive amount of ground-truth data which was difficult to acquire. In this work, we propose an efficient approach to produce a significant amount of realistic training data and introduce a novel recurrent network architecture to deblur frames, which can efficiently handle arbitrary spatial and temporal input sizes.

****

Joint Adaptive Sparsity and Low-Rankness on the Fly: An Online Tensor Reconstruction Scheme for Video Denoising

Bihan Wen, Yanjun Li, Luke Pfister, Yoram Bresler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 241-250

Recent works on adaptive sparse and low-rank signal modeling have demonstrated their usefulness, especially in image/video processing applications. While a patch-based sparse model imposes local structure, low-rankness of the grouped patches exploits non-local correlation. Applying either approach alone usually limits performance in various low-level vision tasks. In this work, we propose a novel video denoising method, based on an online tensor reconstruction scheme with a joint adaptive sparse and low-rank model, dubbed SALT. An efficient and unsupervised online unitary sparsifying transform learning method is introduced to impose adaptive sparsity on the fly. We develop an efficient 3D spatio-temporal data reconstruction framework based on the proposed online learning method, which exhibits low latency and can potentially handle streaming videos. To the best of our knowledge, this is the first work that combines adaptive sparsity and low-rankness for video denoising, and the first work of solving the proposed problem in an online fashion. We demonstrate video denoising results over commonly used videos from public datasets. Numerical experiments show that the proposed video denoising method outperforms competing methods.

****

Learning to Super-Resolve Blurry Face and Text Images

Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 251-260

We present an algorithm to directly restore a clear high-resolution image from a blurry low-resolution input. This problem is highly ill-posed and the basic assumptions for existing super-resolution methods (requiring clear input) and deblurring methods (requiring high-resolution input) no longer hold. We focus on face and text images and adopt a generative adversarial network (GAN) to learn a category-specific prior to solve this problem. However, the basic GAN formulation does not generate realistic high-resolution images. In this work, we introduce novel training losses that help recover fine details. We also present a multi-class GAN that can process multi-class image restoration tasks, i.e., face and text images, using a single generator network. Extensive experiments demonstrate that our method performs favorably against the state-of-the-art methods on both synthetic and real-world images at a lower computational cost.

****

Video Frame Interpolation via Adaptive Separable Convolution

Simon Niklaus, Long Mai, Feng Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 261-270

Standard video frame interpolation methods first estimate optical flow between input frames and then synthesize an intermediate frame guided by motion. Recent approaches merge these two steps into a single convolution process by convolving input frames with spatially adaptive kernels that account for motion and re-sampling simultaneously. These methods require large kernels to handle large motion, which limits the number of pixels whose kernels can be estimated at once due to the large memory demand. To address this problem, this paper formulates frame interpolation as local separable convolution over input frames using pairs of 1D kernels. Compared to regular 2D kernels, the 1D kernels require significantly fewer parameters to be estimated. Our method develops a deep fully convolutional neural network that takes two input frames and estimates pairs of 1D kernels for all pixels simultaneously. Since our method is able to estimate kernels and synthesizes the whole video frame at once, it allows for the incorporation of perceptual loss to train the neural network to produce visually pleasing frames. This deep neural network is trained end-to-end using widely available video data without any human annotation. Both qualitative and quantitative experiments show that our method provides a practical solution to high-quality video frame interpolation.

****

Deep Occlusion Reasoning for Multi-Camera Multi-Target Detection

Pierre Baque, Francois Fleuret, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 271-279

People detection in 2D images has improved greatly in recent years. However, comparatively little of this progress has percolated into multi-camera multi-people tracking algorithms, whose performance still degrades severely when scenes become very crowded. In this work, we introduce a new architecture that combines Convolutional Neural Nets and Conditional Random Fields to explicitly resolve ambiguities. One of its key ingredients are high-order CRF terms that model potential occlusions and give our approach its robustness even when many people are present. Our model is trained end-to-end and we show that it outperforms several state-of-the-art algorithms on challenging scenes.

****

Encouraging LSTMs to Anticipate Actions Very Early

Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, Lars Andersson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 280-289

In contrast to the widely studied problem of recognizing an action given a complete sequence, action anticipation aims to identify the action from only partially available videos. As such, it is therefore key to the success of computer vision applications requiring to react as early as possible, such as autonomous navigation. In this paper, we propose a new action anticipation method that achieves high prediction accuracy even in the presence of a very small percentage of a video sequence. To this end, we develop a multi-stage LSTM architecture that leverages context-aware and action-aware features, and introduce a novel loss function that encourages the model to predict the correct class as early as possible. Our experiments on standard benchmark datasets evidence the benefits of our approach; We outperform the state-of-the-art action anticipation methods for early prediction by a relative increase in accuracy of 22.0% on JHMDB-21, 14.0% on UT-Interaction and 49.9% on UCF-101.

****

PathTrack: Fast Trajectory Annotation With Path Supervision

Santiago Manen, Michael Gygli, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 290-299

Progress in Multiple Object Tracking (MOT) has been limited by the size of the available datasets. We present an efficient framework to annotate trajectories and use it to produce a MOT dataset of unprecedented size. A novel path supervision paradigm lets the annotator loosely track the object with a cursor while watching the video. This results in a path annotation for each object in the sequence. These path annotations, together with object detections, are fed into a two-step optimization to produce full bounding-box trajectories. Our experiments on existing datasets prove that our framework produces more accurate annotations than the state of the art and this in a fraction of the time. We further validate our approach by generating the PathTrack dataset, with more than 15,000 person trajectories in 720 sequences. We believe tracking approaches can benefit from a larger dataset like this one, just as was the case in object recognition. We show its potential by using it to re-train an off-the-shelf person matching network, originally trained on the MOT15 dataset, almost halving the misclassification rate. Additionally, training on our data consistently improves tracking results, both on our dataset and on MOT15. In the latter, where we improve the top-performing tracker (NOMT) dropping the number of ID Switches by 18% and fragments by 5%.

****

Tracking the Untrackable: Learning to Track Multiple Cues With Long-Term Dependencies

Amir Sadeghian, Alexandre Alahi, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 300-311

The majority of existing solutions to the Multi-Target Tracking (MTT) problem do not combine cues over a long period of time in a coherent fashion. In this paper, we present an online method that encodes long-term temporal dependencies across multiple cues. One key challenge of tracking methods is to accurately track occluded targets or those which share similar appearance properties with surrounding objects. To address this challenge, we present a structure of Recurrent Neural Networks (RNN) that jointly reasons on multiple cues over a temporal window. Our method allows to correct data association errors and recover observations from occluded states. We demonstrate the robustness of our data-driven approach by tracking multiple targets using their appearance, motion, and even interactions. Our method outperforms previous works on multiple publicly available datasets including the challenging MOT benchmark.

****

MirrorFlow: Exploiting Symmetries in Joint Optical Flow and Occlusion Estimation

Junhwa Hur, Stefan Roth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 312-321

Optical flow estimation is one of the most studied problems in computer vision, yet recent benchmark datasets continue to reveal problem areas of today's approaches. Occlusions have remained one of the key challenges. In this paper, we propose a symmetric optical flow method to address the well-known chicken-and-egg relation between optical flow and occlusions. In contrast to many state-of-the-art methods that consider occlusions as outliers, possibly filtered out during post-processing, we highlight the importance of joint occlusion reasoning in the optimization and show how to utilize occlusion as an important cue for estimating optical flow. The key feature of our model is to fully exploit the symmetry properties that characterize optical flow and occlusions in the two consecutive images. Specifically through utilizing forward-backward consistency and occlusion-disocclusion symmetry in the energy, our model jointly estimates optical flow in both forward and backward direction, as well as consistent occlusion maps in both views. We demonstrate significant performance benefits on standard benchmarks, especially from the occlusion-disocclusion symmetry. On the challenging KITTI dataset we report the most accurate two-frame results to date.

****

Tracking as Online Decision-Making: Learning a Policy From Streaming Videos With Reinforcement Learning

James Supancic,III, Deva Ramanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 322-331

We formulate tracking as an online decision-making process, where a tracking agent must follow an object despite ambiguous image frames and a limited computational budget. Crucially, the agent must decide where to look in the upcoming frames, when to reinitialize because it believes the target has been lost, and when to update its appearance model for the tracked object. Such decisions are typically made heuristically. Instead, we propose to learn an optimal decision-making policy by formulating tracking as a partially observable decision-making process (POMDP). We learn policies with deep reinforcement learning algorithms that need supervision (a reward signal) only when the track has gone awry. We demonstrate that sparse rewards allow us to quickly train on massive datasets, several orders of magnitude more than past work. Interestingly, by treating the data source of Internet videos as unlimited streams, we both learn and evaluate our trackers in a single, unified computational stream.

****

Non-Convex Rank/Sparsity Regularization and Local Minima

Carl Olsson, Marcus Carlsson, Fredrik Andersson, Viktor Larsson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 332-340

This paper considers the problem of recovering either a low rank matrix or a sparse vector from observations of linear combinations of the vector or matrix elements. Recent methods replace the non-convex regularization with l1 or nuclear norm relaxations. It is well known that this approach recovers near optimal solutions if a so called restricted isometry property (RIP) holds. On the other hand it also has a shrinking bias which can degrade the solution. In this paper we study an alternative non-convex regularization term that does not suffer from this bias. Our main theoretical results show that if a RIP holds then the stationary points are often well separated, in the sense that their differences must be of high cardinality/rank. Thus, with a suitable initial solution the approach is unlikely to fall into a bad local minimum. Our numerical tests show that the approach is likely to converge to a better solution than standard l1/nuclear-norm relaxation even when starting from trivial initializations. In many cases our results can also be used to verify global optimality of our method.

****

A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework

Weixin Luo, Wen Liu, Shenghua Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 341-349

Motivated by the capability of sparse coding based anomaly detection, we propose a Temporally-coherent Sparse Coding (TSC) where we enforce similar neighbouring frames be encoded with similar reconstruction coefficients. Then we map the TSC with a special type of stacked Recurrent Neural Network (sRNN). By taking advantage sRNN in learning all parameters simultaneously, the nontrivial hyper-parameter selection to TSC can be avoided, meanwhile with a shallow sRNN, the reconstruction coefficients can be inferred within a forward pass, which reduces the computational cost for learning sparse coefficients. The contributions of this paper are two-fold: i) We propose a TSC, which can be mapped to a sRNN which facilitates the parameter optimization and accelerates the anomaly prediction. ii) We build a very large dataset which is even larger than the summation of all existing dataset for anomaly detection in terms of both the volume of data and the diversity of scenes. Extensive experiments on both a toy dataset and real datasets demonstrate that our TSC based and sRNN based method consistently outperform existing methods, which validates the effectiveness of our method.

****

HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis

Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 350-359

Pedestrian analysis plays a vital role in intelligent video surveillance and is a key component for security-centric computer vision systems. Despite that the convolutional neural networks are remarkable in learning discriminative features from images, the learning of comprehensive features of pedestrians for fine-grained tasks remains an open problem. In this study, we propose a new attention-based deep neural network, named as HydraPlus-Net (HP-net), that multi-directionally feeds the multi-level attention maps to different feature layers. The attentive deep features learned from the proposed HP-net bring unique advantages: (1) the model is capable of capturing multiple attentions from low-level to semantic-level, and (2) it explores the multi-scale selectiveness of attentive features to enrich the final feature representations for a pedestrian image. We demonstrate the effectiveness and generality of the proposed HP-net for pedestrian analysis on two tasks, i.e. pedestrian attribute recognition and person re-identification. Intensive experimental results have been provided to prove that the HP-net outperforms the state-of-the-art methods on various datasets.

****

No Fuss Distance Metric Learning Using Proxies

Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, Saurabh

Singh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 360-368

We address the problem of distance metric learning (DML), defined as learning a distance consistent with a notion of semantic similarity. Traditionally, for this problem supervision is expressed in the form of sets of points that follow an ordinal relationship -- an anchor point x is similar to a set of positive points Y, and dissimilar to a set of negative points Z, and a loss defined over these distances is minimized. While the specifics of the optimization differ, in this work we collectively call this type of supervision Triplets and all methods that follow this pattern Triplet-Based methods. These methods are challenging to optimize. A main issue is the need for finding informative triplets, which is usually achieved by a variety of tricks such as increasing the batch size, hard or semi-hard triplet mining, etc. Even with these tricks, the convergence rate of such methods is slow. In this paper we propose to optimize the triplet loss on a different space of triplets, consisting of an anchor data point and similar and dissimilar proxy points which are learned as well. These proxies approximate the original data points, so that a triplet loss over the proxies is a tight upper bound of the original loss. This proxy-based loss is empirically better behaved. As a result, the proxy-loss improves on state-of-art results for three standard zero-shot learning datasets, by up to 15% points, while converging three times as fast as other triplet-based losses.

****

Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation

Matteo Ruggero Ronchi, Pietro Perona; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 369-378

We propose a new method to analyze the impact of errors in algorithms for multi-instance pose estimation and a principled benchmark that can be used to compare them. We define and characterize three classes of errors - localization, scoring, and background - study how they are influenced by instance attributes and their impact on an algorithm's performance. Our technique is applied to compare the two leading methods for human pose estimation on the COCO Dataset, measure the sensitivity of pose estimation with respect to instance size, type and number of visible keypoints, clutter due to multiple instances, and the relative score of instances. The performance of algorithms, and the types of error they make, are highly dependent on all these variables, but mostly on the number of keypoints and the clutter. The analysis and software tools we propose offer a novel and insightful approach for understanding the behavior of pose estimation algorithms and an effective method for measuring their strengths and weaknesses.

****

Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-Identification

Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 379-387

In this paper, we tackle the vehicle Re-identification (ReID) problem which is of great importance in urban surveillance and can be used for multiple applications. In our vehicle ReID framework, an orientation invariant feature embedding module and a spatial-temporal regularization module are proposed. With orientation invariant feature embedding, local region features of different orientations can be extracted based on 20 key point locations and can be well aligned and combined. With spatial-temporal regularization, the log-normal distribution is adopted to model the spatial-temporal constraints and the retrieval results can be refined. Experiments are conducted on public vehicle ReID datasets and our proposed method achieves state-of-the-art performance. Investigations of the proposed framework is conducted, including the landmark regressor and comparisons with attention mechanism. Both the orientation invariant feature embedding and the spatio-temporal regularization achieve considerable improvements.

****

Fashion Forward: Forecasting Visual Style in Fashion

Ziad Al-Halah, Rainer Stiefelhagen, Kristen Grauman; Proceedings of the IEEE Int ernational Conference on Computer Vision (ICCV), 2017, pp. 388-397

What is the future of fashion? Tackling this question from a data-driven vision perspective, we propose to forecast visual style trends before they occur. We introduce the first approach to predict the future popularity of styles discovered from fashion images in an unsupervised manner. Using these styles as a basis, we train a forecasting model to represent their trends over time. The resulting model can hypothesize new mixtures of styles that will become popular in the future, discover style dynamics (trendy vs. classic), and name the key visual attributes that will dominate tomorrow's fashion. We demonstrate our idea applied to three datasets encapsulating 80,000 fashion products sold across six years on Amazon. Results indicate that fashion forecasting benefits greatly from visual analysis, much more than textual or meta-data cues surrounding products.

****

Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, Yichen Wei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 398-407

In this paper, we study the task of 3D human pose estimation in the wild. This task is challenging due to lack of training data, as existing datasets are either in the wild images with 2D pose or in the lab images with 3D pose. We propose a weakly-supervised transfer learning method that uses mixed 2D and 3D labels in a unified deep neutral network that presents two-stage cascaded structure. Our network augments a state-of-the-art 2D pose estimation sub-network with a 3D depth regression sub-network. Unlike previous two stage approaches that train the two sub-networks sequentially and separately, our training is end-to-end and fully exploits the correlation between the 2D pose and depth estimation sub-tasks. The deep features are better learnt through shared representations. In doing so, the 3D pose labels in controlled lab environments are transferred to in the wild images. In addition, we introduce a 3D geometric constraint to regularize the 3D pose prediction, which is effective in the absence of ground truth depth labels. Our method achieves competitive results on both 2D and 3D benchmarks.

****

Flow-Guided Feature Aggregation for Video Object Detection

Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, Yichen Wei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 408-417

Extending state-of-the-art object detectors from image to video is challenging. The accuracy of detection suffers from degenerated object appearances in videos, e.g., motion blur, video defocus, rare poses, etc. Existing work attempts to exploit temporal information on box level, but such methods are not trained end-to-end. We present flow-guided feature aggregation, an accurate and end-to-end learning framework for video object detection. It leverages temporal coherence on feature level instead. It improves the per-frame features by aggregation of nearby features along the motion paths, and thus improves the video recognition accuracy. Our method significantly improves upon strong single-frame baselines in ImageNet VID, especially for more challenging fast moving objects. Our framework is principled, and on par with the best engineered systems winning the ImageNet VID challenges 2016, without additional bells-and-whistles.

****

Reasoning About Fine-Grained Attribute Phrases Using Reference Games

Jong-Chyi Su, Chenyun Wu, Huaizu Jiang, Subhransu Maji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 418-427

We present a framework for learning to describe fine-grained visual differences between instances using attribute phrases. Attribute phrases capture distinguishing aspects of an object (e.g., "propeller on the nose" or "door near the wing" for airplanes) in a compositional manner. Instances within a category can be described by a set of these phrases and collectively they span the space of semantic attributes for a category. We collect a large dataset of such phrases by asking annotators to describe several visual differences between a pair of instances within a category. We then learn to describe and ground these phrases to images in the context of a *reference game* between a speaker and a listener. The goal of a speaker is to describe attributes of an image that allows the listener to c orrectly identify it within a pair. Data collected in a pairwise manner improves the ability of the speaker to generate, and the ability of the listener to interpret visual descriptions. Moreover, due to the compositionality of attribute phrases, the trained listeners can interpret descriptions not seen during training for image retrieval, and the speakers can generate attribute-based explanations for differences between previously unseen categories. We also show that embedding an image into the semantic space of attribute phrases derived from listeners offers 20% improvement in accuracy over existing attribute-based representations on the FGVC-aircraft dataset.

****

DeNet: Scalable Real-Time Object Detection With Directed Sparse Sampling

Lachlan Tychsen-Smith, Lars Petersson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 428-436

We define the object detection from imagery problem as estimating a very large but extremely sparse bounding box dependent probability distribution. Subsequently we identify a sparse distribution estimation scheme, Directed Sparse Sampling, and employ it in a single end-to-end CNN based detection model. This methodology extends and formalizes previous state-of-the-art detection models with an additional emphasis on high evaluation rates and reduced manual engineering. We introduce two novelties, a corner based region-of-interest estimator and a deconvolution based CNN model. The resulting model is scene adaptive, does not require manually defined reference bounding boxes and produces highly competitive results on MSCOCO, Pascal VOC 2007 and Pascal VOC 2012 with real-time evaluation rates. Further analysis suggests our model performs particularly well when finegrained object localization is desirable. We argue that this advantage stems from the significantly larger set of available regions-of-interest relative to other methods. Source-code is available from: https://github.com/lachlants/denet

****

MIHash: Online Hashing With Mutual Information

Fatih Cakir, Kun He, Sarah Adel Bargal, Stan Sclaroff; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 437-445

Learning-based hashing methods are widely used for nearest neighbor retrieval, and recently, online hashing methods have demonstrated good performance-complexity trade-offs by learning hash functions from streaming data. In this paper, we first address a key challenge for online hashing: the binary codes for indexed data must be recomputed to keep pace with updates to the hash functions. We propose an efficient quality measure for hash functions, based on an information-theoretic quantity, mutual information, and use it successfully as a criterion to eliminate unnecessary hash table updates. Next, we also show how to optimize the mutual information objective using stochastic gradient descent. We thus develop a novel hashing method, MIHash, that can be used in both online and batch settings. Experiments on image retrieval benchmarks (including a 2.5M image dataset) confirm the effectiveness of our formulation, both in reducing hash table recomputations and in learning high-quality hash functions.

****

SafetyNet: Detecting and Rejecting Adversarial Examples Robustly

Jiajun Lu, Theerasit Issaranon, David Forsyth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 446-454

We describe a method to produce a network where current methods such as DeepFool have great difficulty producing adversarial samples. Our construction suggests some insights into how deep networks work. We provide a reasonable analyses that our construction is difficult to defeat, and show experimentally that our method is hard to defeat with both Type I and Type II attacks using several standard networks and datasets. This SafetyNet architecture is used to an important and novel application SceneProof, which can reliably detect whether an image is a picture of a real scene or not. SceneProof applies to images captured with depth maps (RGBD images) and checks if a pair of image and depth map is consistent. It relies on the relative difficulty of producing naturalistic depth maps for images in post processing. We demonstrate that our SafetyNet is robust to adversarial examples built from currently known attacking approaches.

****

Recurrent Models for Situation Recognition

Arun Mallya, Svetlana Lazebnik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 455-463

This work proposes Recurrent Neural Network (RNN) models to predict structured 'image situations' -- actions and noun entities fulfilling semantic roles related to the action. In contrast to prior work relying on Conditional Random Fields (CRFs), we use a specialized action prediction network followed by an RNN for noun prediction. Our system obtains state-of-the-art accuracy on the challenging recent imSitu dataset, beating CRF-based models, including ones trained with additional data. Further, we show that specialized features learned from situation prediction can be transferred to the task of image captioning to more accurately describe human-object interactions.

****

Multi-Label Image Recognition by Recurrently Discovering Attentional Regions

Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, Liang Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 464-472

This paper proposes a novel deep architecture to address multi-label image recognition, a fundamental and practical task towards general visual understanding. Current solutions for this task usually rely on an extra step of extracting hypothesis regions (i.e., region proposals), resulting in redundant computation and sub-optimal performance. In this work, we achieve the interpretable and contextualized multi-label image classification by developing a recurrent memorized-attention module. This module consists of two alternately performed components: i) a spatial transformer layer to locate attentional regions from the convolutional feature maps in a region-proposal-free way and ii) a LSTM (Long-Short Term Memory) sub-network to sequentially predict semantic labeling scores on the located regions while capturing the global dependencies of these regions. The LSTM also output the parameters for computing the spatial transformer. On large-scale benchmarks of multi-label image classification (e.g., MS-COCO and PASCAL VOC 07), our approach demonstrates superior performances over other existing state-of-the-arts in both accuracy and efficiency.

****

Deep Determinantal Point Process for Large-Scale Multi-Label Classification

Pengtao Xie, Ruslan Salakhutdinov, Luntian Mou, Eric P. Xing; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 473-482

We study large-scale multi-label classification (MLC) on two recently released datasets: Youtube-8M and Open Images that contain millions of data instances and thousands of classes. The unprecedented problem scale poses great challenges for MLC. First, finding out the correct label subset out of exponentially many choices incurs substantial ambiguity and uncertainty. Second, the large data-size and class-size entail considerable computational cost. To address the first challenge, we investigate two strategies: capturing label-correlations from the training data and incorporating label co-occurrence relations obtained from external knowledge, which effectively eliminate semantically inconsistent labels and provide contextual clues to differentiate visually ambiguous labels. Specifically, we propose a Deep Determinantal Point Process (DDPP) model which seamlessly integrates a DPP with deep neural networks (DNNs) and supports end-to-end multi-label learning and deep representation learning. The DPP is able to capture label-correlations of any order with a polynomial computational cost, while the DNNs learn hierarchical features of images/videos and capture the dependency between input data and labels. To incorporate external knowledge about label co-occurrence relations, we impose a relational regularization over the kernel matrix in DDPP. To address the second challenge, we study an efficient low-rank kernel learning algorithm based on inducing point methods. Experiments on the two datasets demonstrate the efficacy and efficiency of the proposed methods.

****

Visual Semantic Planning Using Deep Successor Representations

Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Gupta, Roozbeh Mottaghi, Ali Farhadi; Proceedings of the IEEE International Conference on C omputer Vision (ICCV), 2017, pp. 483-492

A crucial capability of real-world intelligent agents is their ability to plan a sequence of actions to achieve their goals in the visual world. In this work, we address the problem of visual semantic planning: the task of predicting a sequence of actions from visual observations that transform a dynamic environment from an initial state to a goal state. Doing so entails knowledge about objects and their affordances, as well as actions and their preconditions and effects. We propose learning these through interacting with a visual and dynamic environment. Our proposed solution involves bootstrapping reinforcement learning with imitation learning. To ensure cross task generalization, we develop a deep predictive model based on successor representations. Our experimental results show near optimal results across a wide range of tasks in the challenging THOR environment. The supplementary video can be accessed at the following link: https://goo.gl/vXsbQP.

****

Neural Person Search Machines

Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 493-501

We investigate the problem of person search in the wild in this work. Instead of comparing the query against all candidate regions generated in a query-blind manner, we propose to recursively shrink the search area from the whole image till achieving precise localization of the target person, by fully exploiting information from the query and contextual cues in every recursive search step. We develop the Neural Person Search Machines (NPSM) to implement such recursive localization for person search. Benefiting from its neural search mechanism, NPSM is able to selectively shrink its focus from a loose region to a tighter one containing the target automatically. In this process, NPSM employs an internal primitive memory component to memorize the query representation which modulates the attention and augments its robustness to other distracting regions. Evaluations on two benchmark datasets, CUHK-SYSU Person Search dataset and PRW dataset, have demonstrated that our method can outperform current state-of-the-arts in both mAP and top-1 evaluation protocols.

****

DualNet: Learn Complementary Features for Image Recognition

Saihui Hou, Xu Liu, Zilei Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 502-510

In this work we propose a novel framework named DualNet aiming at learning more accurate representation for image recognition. Here two parallel neural networks are coordinated to learn complementary features and thus a wider network is constructed. Specifically, we logically divide an end-to-end deep convolutional neural network into two functional parts, i.e., feature extractor and image classifier. The extractors of two subnetworks are placed side by side, which exactly form the feature extractor of DualNet. Then the two-stream features are aggregated to the final classifier for overall classification, while two auxiliary classifiers are appended behind the feature extractor of each subnetwork to make the separately learned features discriminative alone. The complementary constraint is imposed by weighting the three classifiers, which is indeed the key of DualNet. The corresponding training strategy is also proposed, consisting of iterative training and joint finetuning, to make the two subnetworks cooperate well with each other. Finally, DualNet based on the well-known CaffeNet, VGGNet, NIN and ResNet are thoroughly investigated and experimentally evaluated on multiple datasets including CIFAR-100, Stanford Dogs and UEC FOOD-100. The results demonstrate that DualNet can really help learn more accurate image representation, and thus result in higher accuracy for recognition. In particular, the performance on CIFAR-100 is state-of-the-art compared to the recent works.

****

Higher-Order Integration of Hierarchical Convolutional Activations for Fine-Grained Visual Categorization

Sijia Cai, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE International Confer ence on Computer Vision (ICCV), 2017, pp. 511-520

The success of fine-grained visual categorization (FGVC) extremely relies on the modeling of appearance and interactions of various semantic parts. This makes FGVC very challenging because: (i) part annotation and detection require expert guidance and are very expensive; (ii) parts are of different sizes; and (iii) the part interactions are complex and of higher-order. To address these issues, we propose an end-to-end framework based on higher-order integration of hierarchical convolutional activations for FGVC. By treating the convolutional activations as local descriptors, hierarchical convolutional activations can serve as a representation of local parts from different scales. A polynomial kernel based predictor is proposed to capture higher-order statistics of convolutional activations for modeling part interaction. To model inter-layer part interactions, we extend polynomial predictor to integrate hierarchical activations via kernel fusion. Our work also provides a new perspective for combining convolutional activations from multiple layers. While hypercolumns simply concatenate maps from different layers, and holistically-nested network uses weighted fusion to combine side-outputs, our approach exploits higher-order intra-layer and inter-layer relations for better integration of hierarchical convolutional features. The proposed framework yields more discriminative representation and achieves competitive results on the widely used FGVC datasets.

****

Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, Min Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 521-530

Impressive image captioning results are achieved in domains with plenty of training image and sentence pairs (e.g., MSCOCO). However, transferring to a target domain with significant domain shifts but no paired training data (referred to as cross-domain image captioning) remains largely unexplored. We propose a novel adversarial training procedure to leverage unpaired data in the target domain. Two critic networks are introduced to guide the captioner, namely domain critic and multi-modal critic. The domain critic assesses whether the generated sentences are indistinguishable from sentences in the target domain. The multi-modal critic assesses whether an image and its generated sentence are a valid pair. During training, the critics and captioner act as adversaries -- captioner aims to generate indistinguishable sentences, whereas critics aim at distinguishing them. The assessment improves the captioner through policy gradient updates. During inference, we further propose a novel critic-based planning method to select high-quality sentences without additional supervision (e.g., tags). To evaluate, we use MSCOCO as the source domain and four other datasets (CUB-200-2011, Oxford-102, TGIF, and Flickr30k) as the target domains. Our method consistently performs well on all datasets. In particular, on CUB-200-2011, we achieve 21.8% CIDEr-D improvement after adaptation. Utilizing critics during inference further gives another 4.5% boost.

****

Attribute Recognition by Joint Recurrent Learning of Context and Correlation

Jingya Wang, Xiatian Zhu, Shaogang Gong, Wei Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 531-540

Recognising semantic pedestrian attributes in surveillance images is a challenging task for computer vision, particularly when the imaging quality is poor with complex background clutter and uncontrolled viewing conditions, and the number of labelled training data is small. In this work, we formulate a Joint Recurrent Learning (JRL) model for exploring attribute context and correlation in order to improve attribute recognition given small sized training data with poor quality images. The JRL model learns jointly pedestrian attribute correlations in a pedestrian image and in particular their sequential ordering dependencies (latent high-order correlation) in an end-to-end encoder/decoder recurrent network. We demonstrate the performance advantage and robustness of the JRL model over a wide range of state-of-the-art deep models for pedestrian attribute recognition, multi-label image classification, and multi-person image annotation on two largest p edestrian attribute benchmarks PETA and RAP.

****

VegFru: A Domain-Specific Dataset for Fine-Grained Visual Categorization

Saihui Hou, Yushan Feng, Zilei Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 541-549

VegFru: A Domain-Specific Dataset for Fine-grained Visual Categorization In this paper, we propose a novel domain-specific dataset named VegFru for fine-grained visual categorization (FGVC). While the existing datasets for FGVC are mainly focused on animal breeds or man-made objects with limited labelled data, VegFru is a larger dataset consisting of vegetables and fruits which are closely associated with the daily life of everyone. Aiming at domestic cooking and food management, VegFru categorizes vegetables and fruits according to their eating characteristics, and each image contains at least one edible part of vegetables or fruits with the same cooking usage. Particularly, all the images are labelled hierarchically. The current version covers vegetables and fruits of 25 upper-level categories and 292 subordinate classes. And it contains more than 160,000 images in total and at least 200 images for each subordinate class. Accompanying the dataset, we also propose an effective framework called HybridNet to exploit the label hierarchy for FGVC. Specifically, multiple granularity features are first extracted by dealing with the hierarchical labels separately. And then they are fused through explicit operation, e.g., Compact Bilinear Pooling, to form a unified representation for the ultimate recognition. The experimental results on the novel VegFru, the public FGVC-Aircraft and CUB-200-2011 indicate that HybridNet achieves one of the top performance on these datasets. The dataset and code are available at https://github.com/hshustc/vegfru.

****

Increasing CNN Robustness to Occlusions by Reducing Filter Support

Elad Osherov, Michael Lindenbaum; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 550-561

Convolutional neural networks (CNNs) provide the current state of the art in visual object classification, but they are far less accurate when classifying partially occluded objects. A straightforward way to improve classification under occlusion conditions is to train the classifier using partially occluded object examples. However, training the network on many combinations of object instances and occlusions may be computationally expensive. This work proposes an alternative approach to increasing the robustness of CNNs to occlusion. We start by studying the effect of partial occlusions on the trained CNN and show, empirically, that training on partially occluded examples reduces the spatial support of the filters. Building upon this finding, we argue that smaller filter support is beneficial for occlusion robustness. We propose a training process that uses a special regularization term that acts to shrink the spatial support of the filters. We consider three possible regularization terms that are based on second central moments, group sparsity, and mutually reweighted L1, respectively. When trained on normal (unoccluded) examples, the resulting classifier is highly robust to occlusions. For large training sets and limited training time, the proposed classifier is even more accurate than standard classifiers trained on occluded object examples.

****

Exploiting Multi-Grain Ranking Constraints for Precisely Searching Visually-Similar Vehicles

Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, Tiejun Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 562-570

Precise search of visually-similar vehicles poses a great challenge in computer vision, which needs to find exactly the same vehicle among a massive vehicles with visually similar appearances for a given query image. In this paper, we model the relationship of vehicle images as multiple grains. Following this, we propose two approaches to alleviate the precise vehicle search problem by exploiting multi-grain ranking constraints. One is Generalized Pairwise Ranking, which generalizes the conventional pairwise from considering only binary similar/dissimilar relations to multiple relations. The other is Multi-Grain based List Ranking, which introduces permutation probability to score a permutation of a multi-grain list, and further optimizes the ranking by the likelihood loss function. We implement the two approaches with multi-attribute classification in a multi-task deep learning framework. To further facilitate the research on precise vehicle search, we also contribute two high-quality and well-annotated vehicle datasets, named VD1 and VD2, which are collected from two different cities with diverse annotated attributes. As two of the largest publicly available precise vehicle search datasets, they contain 1,097,649 and 807,260 vehicle images respectively. Experimental results show that our approaches achieve the state-of-the-art performance on both datasets.

****

Recurrent Scale Approximation for Object Detection in CNN

Yu Liu, Hongyang Li, Junjie Yan, Fangyin Wei, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 571-579

Since convolutional neural network (CNN) lacks an inherent mechanism to handle large scale variations, we always need to compute feature maps multiple times for multi-scale object detection, which has the bottleneck of computational cost in practice. To address this, we devise a recurrent scale approximation (RSA) to compute feature map once only, and only through this map can we approximate the rest maps on other levels. At the core of RSA is the recursive rolling out mechanism: given an initial map on a particular scale, it generates the prediction on a smaller scale that is half the size of input. To further increase efficiency and accuracy, we (a): design a scale-forecast network to globally predict potential scales in the image since there is no need to compute maps on all levels of the pyramid. (b): propose a landmark retracing network (LRN) to retrace back locations of the regressed landmarks and generate a confidence score for each landmark; LRN can effectively alleviate false positives due to the accumulated error in RSA. The whole system could be trained end-to-end in a unified CNN framework. Experiments demonstrate that our proposed algorithm is superior against state-of-the-arts on face detection benchmarks and achieves comparable results for generic proposal generation. The source code of RSA is available at github.com/sciencefans/RSA-for-object-detection.

****

Embedding 3D Geometric Features for Rigid Object Part Segmentation

Yafei Song, Xiaowu Chen, Jia Li, Qinping Zhao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 580-588

Object part segmentation is a challenging and fundamental problem in computer vision. Its difficulties may be caused by the varying viewpoints, poses, and topological structures, which can be attributed to an essential reason, i.e., a specific object is a 3D model rather than a 2D figure. Therefore, we conjecture that not only 2D appearance features but also 3D geometric features could be helpful. With this in mind, we propose a 2-stream FCN. One stream, named AppNet, is to extract 2D appearance features from the input image. The other stream, named GeoNet, is to extract 3D geometric features. However, the problem is that the input is just an image. To this end, we design a 2D-convolution based CNN structure to extract 3D geometric features from 3D volume, which is named VolNet. Then a teacher-student strategy is adopted and VolNet teaches GeoNet how to extract 3D geometric features from an image. To perform this teaching process, we synthesize training data using 3D models. Each training sample consists of an image and its corresponding volume. A perspective voxelization algorithm is further proposed to align them. Experimental results verify our conjecture and the effectiveness of both the proposed 2-stream CNN and VolNet.

****

Towards Context-Aware Interaction Recognition for Visual Relationship Detection

Bohan Zhuang, Lingqiao Liu, Chunhua Shen, Ian Reid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 589-598

Recognizing how objects interact with each other is a crucial task in visual recognition. If we define the context of the interaction to be the objects involved, then most current methods can be categorized as either: (i) training a single classifier on the combination of the interaction and its context; or (ii) aiming to recognize the interaction independently of its explicit context. Both methods suffer limitations: the former scales poorly with the number of combinations and fails to generalize to unseen combinations, while the latter often leads to poor interaction recognition performance due to the difficulty of designing a context-independent interaction classifier. To mitigate those drawbacks, this paper proposes an alternative, context-aware interaction recognition framework. The key to our method is to explicitly construct an interaction classifier which combines the context, and the interaction. The context is encoded via word2vec into a semantic space, and is used to derive a classification result for the interaction. The proposed method still builds one classifier for one interaction (as per type (ii) above), but the classifier built is adaptive to context via weights which are context dependent. The benefit of using the semantic space is that it naturally leads to zero-shot generalizations in which semantically similar contexts (subject-object pairs) can be recognized as suitable contexts for an interaction, even if they were not observed in the training set. Our method also scales with the number of interaction-context pairs since our model parameters do not increase with the number of interactions. Thus our method avoids the limitation of both approaches. We demonstrate experimentally that the proposed framework leads to improved performance for all investigated interaction representations and datasets.

****

When Unsupervised Domain Adaptation Meets Tensor Representations
Hao Lu, Lei Zhang, Zhiguo Cao, Wei Wei, Ke Xian, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 599-608

Domain adaption (DA) allows machine learning methods trained on data sampled from one distribution to be applied to data sampled from another. It is thus of great practical importance to the application of such methods. Despite the fact that tensor representations are widely used in Computer Vision to capture multi-linear relationships that affect the data, most existing DA methods are applicable to vectors only. This renders them incapable of reflecting and preserving important structure in many problems. We thus propose here a learning-based method to adapt the source and target tensor representations directly, without vectorization. In particular, a set of alignment matrices is introduced to align the tensor representations from both domains into the invariant tensor subspace. These alignment matrices and the tensor subspace are modeled as a joint optimization problem and can be learned adaptively from the data using the proposed alternative minimization scheme. Extensive experiments show that our approach is capable of preserving the discriminative power of the source domain, of resisting the effects of label noise, and works effectively for small sample sizes, and even one-shot DA. We show that our method outperforms the state-of-the-art on the task of cross-domain visual recognition in both efficacy and efficiency, and particularly that it outperforms all comparators when applied to DA of the convolutional activations of deep convolutional networks.

****

Look, Listen and Learn
Relja Arandjelovic, Andrew Zisserman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 609-617

We consider the question: what can be learnt by looking at and listening to a large number of unlabelled videos? There is a valuable, but so far untapped, source of information contained in the video itself -- the correspondence between the visual and the audio streams, and we introduce a novel "Audio-Visual Correspondence" learning task that makes use of this. Training visual and audio networks from scratch, without any additional supervision other than the raw unconstrained videos themselves, is shown to successfully solve this task, and, more interestingly, result in good visual and audio representations. These features set the new state-of-the-art on two sound classification benchmarks, and perform on par with the state-of-the-art self-supervised approaches on ImageNet classification. We also demonstrate that the network is able to localize objects in both modalit ies, as well as perform fine-grained recognition tasks.

****

Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626

We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. Our approach - Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say logits for 'dog' or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (e.g. VGG), (2) CNNs used for structured outputs (e.g. captioning), (3) CNNs used in tasks with multi-modal inputs (e.g. VQA) or reinforcement learning, and needs no architectural changes or re-training. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures. In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are more faithful to the underlying model, and (d) help achieve model generalization by identifying dataset bias. For image captioning and VQA, our visualizations show that even non-attention based models can localize inputs. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions. Our code is available at https://github.com/ramprs/grad-cam/ along with a demo on CloudCV [2] 1 and video at youtu.be/COjUB9Izk6E.

****

Image-Based Localization Using LSTMs for Structured Feature Correlation

Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 627-637

In this work we propose a new CNN+LSTM architecture for camera pose regression for indoor and outdoor scenes. CNNs allow us to learn suitable feature representations for localization that are robust against motion blur and illumination changes. We make use of LSTM units on the CNN output, which play the role of a structured dimensionality reduction on the feature vector, leading to drastic improvements in localization performance. We provide extensive quantitative comparison of CNN-based and SIFT-based localization methods, showing the weaknesses and strengths of each. Furthermore, we present a new large-scale indoor dataset with accurate ground truth from a laser scanner. Experimental results on both indoor and outdoor public datasets show our method outperforms existing deep architectures, and can localize images in hard conditions, e.g., in the presence of mostly textureless surfaces, where classic SIFT-based methods fail.

****

Personalized Image Aesthetics

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, David J. Foran; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 638-647

Automatic image aesthetics rating has received a growing interest with the recent breakthrough in deep learning. Although many studies exist for learning a generic or universal aesthetics model, investigation of aesthetics models incorporating individual user's preference is quite limited. We address this personalized aesthetics problem by showing that individual's aesthetic preferences exhibit strong correlations with content and aesthetic attributes, and hence the deviation of individual's perception from generic image aesthetics is predictable. To acc ommodate our study, we first collect two distinct datasets, a large image dataset from Flickr and annotated by Amazon Mechanical Turk, and a small dataset of real personal albums rated by owners. We then propose a new approach to personalized aesthetics learning that can be trained even with a small set of annotated images from a user. The approach is based on a residual-based model adaptation scheme which learns an offset to compensate for the generic aesthetics score. Finally, we introduce an active learning algorithm to optimize personalized aesthetics prediction for real-world application scenarios. Experiments demonstrate that our approach can effectively learn personalized aesthetics preferences, and outperforms existing methods on quantitative comparisons.

****

Predicting Deeper Into the Future of Semantic Segmentation
Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, Yann LeCun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 648-657

The ability to predict and therefore to anticipate the future is an important attribute of intelligence. It is also of utmost importance in real-time systems, e.g . in robotics or autonomous driving, which depend on visual scene understanding for decision making. While prediction of the raw RGB pixel values in future video frames has been studied in previous work, here we introduce the novel task of predicting semantic segmentations of future frames. Given a sequence of video frames, our goal is to predict segmentation maps of not yet observed video frames that lie up to a second or further in the future. We develop an autoregressive convolutional neural network that learns to iteratively generate multiple frames. Our results on the Cityscapes dataset show that directly predicting future segmentations is substantially better than predicting and then segmenting future RGB frames. Prediction results up to half a second in the future are visually convincing and are much more accurate than those of a baseline based on warping semantic segmentations using optical flow.

****

Coordinating Filters for Faster Deep Neural Networks
Wei Wen, Cong Xu, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 658-666

Very large-scale Deep Neural Networks (DNNs) have achieved remarkable successes in a large variety of computer vision tasks. However, the high computation intensity of DNNs makes it challenging to deploy these models on resource-limited systems. Some studies used low-rank approaches that approximate the filters by low-rank basis to accelerate the testing. Those works directly decomposed the pre-trained DNNs by Low-Rank Approximations (LRA). How to train DNNs toward lower-rank space for more efficient DNNs, however, remains as an open area. To solve the issue, in this work, we propose Force Regularization, which uses attractive forces to enforce filters so as to coordinate more weight information into lower-rank space. We mathematically and empirically verify that after applying our technique, standard LRA methods can reconstruct filters using much lower basis and thus result in faster DNNs. The effectiveness of our approach is comprehensively evaluated in ResNets, AlexNet, and GoogLeNet. In AlexNet, for example, Force Regularization gains 2x speedup on modern GPU without accuracy loss and 4.05x speedup on CPU by paying small accuracy degradation. Moreover, Force Regularization better initializes the low-rank DNNs such that the fine-tuning can converge faster toward higher accuracy. The obtained lower-rank DNNs can be further sparsified, proving that Force Regularization can be integrated with state-of-the-art sparsity-based acceleration methods.

****

Unsupervised Representation Learning by Sorting Sequences
Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 667-676

We present an unsupervised representation learning approach using videos without semantic labels. We leverage the temporal coherence as a supervisory signal by formulating representation learning as a sequence sorting task. We take temporally shuffled frames (i.e. in non-chronological order) as inputs and train a convo lutional neural network to sort the shuffled sequences. Similar to comparison-based sorting algorithms, we propose to extract features from all frame pairs and aggregate them to predict the correct order. As sorting shuffled image sequence requires an understanding of the statistical temporal structure of images, training with such a proxy task allows us to learn rich and generalizable visual representation. We validate the effectiveness of the learned representation using our method as pre-training on high-level recognition problems. The experimental results show that our method compares favorably against state-of-the-art methods on action recognition, image classification and object detection tasks.

****

A Read-Write Memory Network for Movie Story Understanding
Seil Na, Sangho Lee, Jisung Kim, Gunhee Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 677-685

We propose a novel memory network model named Read-Write Memory Network (RWMN) to perform question and answering tasks for large-scale, multimodal movie story understanding. The key focus of our RWMN model is to design the read network and the write network that consist of multiple convolutional layers, which enable memory read and write operations to have high capacity and flexibility. While existing memory-augmented network models treat each memory slot as an independent block, our use of multi-layered CNNs allows the model to read and write sequential memory cells as chunks, which is more reasonable to represent a sequential story because adjacent memory blocks often have strong correlations. For evaluation, we apply our model to all the six tasks of the MovieQA benchmark, and achieve the best accuracies on several tasks, especially on the visual QA task. Our model shows a potential to better understand not only the content in the story, but also more abstract information, such as relationships between characters and the reasons for their actions.

****

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow
Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 686-695

This paper proposes an end-to-end trainable network, SegFlow, for simultaneously predicting pixel-wise object segmentation and optical flow in videos. The proposed SegFlow has two branches where useful information of object segmentation and optical flow is propagated bidirectionally in a unified framework. The segmentation branch is based on a fully convolutional network, which has been proved effective in image segmentation task, and the optical flow branch takes advantage of the FlowNet model. The unified framework is trained iteratively offline to learn a generic notion, and fine tuned online for specific objects. Extensive experiments on both the video object segmentation and optical flow datasets demonstrate that introducing optical flow improves the performance of segmentation and vice versa, against the state-of-the-art algorithms.

****

Unsupervised Action Discovery and Localization in Videos
Khurram Soomro, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 696-705

This paper is the first to address the problem of unsupervised action localization in videos. Given unlabeled data without bounding box annotations, we propose a novel approach that: 1) Discovers action class labels and 2) Spatio-temporally localizes actions in videos. It begins by computing local video features to apply spectral clustering on a set of unlabeled training videos. For each cluster of videos, an undirected graph is constructed to extract a dominant set, which are known for high internal homogeneity and inhomogeneity between vertices outside it. Next, a discriminative clustering approach is applied, by training a classifier for each cluster, to iteratively select videos from the non dominant set and obtain complete video action classes. Once classes are discovered, training videos within each cluster are selected to perform automatic spatio-temporal annotations, by first oversegmenting videos in each discovered class into supervoxels and constructing a directed graph to apply a variant of knapsack problem with temporal constraints. Knapsack optimization jointly collects a subset of supervox els, by enforcing the annotated action to be spatio-temporally connected and its volume to be the size of an actor. These annotations are used to train SVM action classifiers. During testing, actions are localized using a similar Knapsack approach, where supervoxels are grouped together and SVM, learned using videos from discovered action classes, is used to recognize these actions. We evaluate our approach on UCF Sports, Sub-JHMDB, JHMDB, THUMOS13 and UCF101 datasets. Our experiments suggest that despite using no action class labels and no bounding box annotations, we are able to get competitive results to the state-of-the-art supervised methods.

****

Dense-Captioning Events in Videos

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 706-715

Most natural videos contain numerous events. For example, in a video of a "man playing a piano", the video might also contain "another man dancing" or "a crowd clapping". We introduce the task of dense-captioning events, which involves both detecting and describing events in a video. We propose a new model that is able to identify all such events in a single pass of the video while simultaneously describing the detected events with natural language. Our model introduces a variant of an existing proposal module that is designed to capture both short as well as long events that span minutes. To capture the dependencies between the events in a video, our model introduces a new captioning module that uses contextual information from past and future events to jointly describe all events. We also introduce ActivityNet Captions, a large-scale benchmark for dense-captioning events. ActivityNet Captions contains 20k videos amounting to 849 video hours with 100k total descriptions, each with it's unique start and end time. Finally, we report performances of our model for dense-captioning events, video retrieval and localization.

****

Learning Long-Term Dependencies for Action Recognition With a Biologically-Inspired Deep Network

Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, Tiejun Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 716-725

Despite a lot of research efforts devoted in recent years, how to efficiently learn long-term dependencies from sequences still remains a pretty challenging task. As one of the key models for sequence learning, recurrent neural network (RNN) and its variants such as long short term memory (LSTM) and gated recurrent unit (GRU) are still not powerful enough in practice. One possible reason is that they have only feedforward connections, which is different from the biological neural system that is typically composed of both feedforward and feedback connections. To address this problem, this paper proposes a biologically-inspired deep network, called shuttleNet. Technologically, the shuttleNet consists of several processors, each of which is a GRU while associated with multiple groups of hidden states. Unlike traditional RNNs, all processors inside shuttleNet are loop connected to mimic the brain's feedforward and feedback connections, in which they are shared across multiple pathways in the loop connection. Attention mechanism is then employed to select the best information flow pathway. Extensive experiments conducted on two benchmark datasets (i.e UCF101 and HMDB51) show that we can beat state-of-the-art methods by simply embedding shuttleNet into a CNN-RNN framework.

****

Compressive Quantization for Fast Object Instance Search in Videos

Tan Yu, Zhenzhen Wang, Junsong Yuan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 726-735

Most of current visual search systems focus on image-to-image (point-to-point) search such as image and object retrieval. Nevertheless, fast image-to-video (point-to-set) search is much less exploited. This paper tackles object instance search in videos, where efficient point-to-set matching is essential. Through jointly optimizing vector quantization and hashing, we propose compressive quantizati on method to compress M object proposals extracted from each video into only k binary codes, where k<< M. Then the similarity between the query object and the whole video can be determined by the Hamming distance between the query's binary code and the video's best-matched binary code. Our compressive quantization not only enables fast search but also significantly reduces the memory cost of storing the video features. Despite the high compression ratio, our proposed compressive quantization still can effectively retrieve small objects in large video datasets. Systematic experiments on three benchmark datasets verify the effectiveness and efficiency of our compressive quantization.

****

Complex Event Detection by Identifying Reliable Shots From Untrimmed Videos

Hehe Fan, Xiaojun Chang, De Cheng, Yi Yang, Dong Xu, Alexander G. Hauptmann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 736-744

The goal of complex event detection is to automatically detect whether an event of interest happens in temporally untrimmed long videos which usually consist of multiple video shots. Observing some video shots in positive (resp. negative) videos are irrelevant (resp. relevant) to the given event class, we formulate this task as a multi-instance learning (MIL) problem by taking each video as a bag and the video shots in each video as instances. To this end, we propose a new MIL method, which simultaneously learns a linear SVM classifier and infers a binary indicator for each instance in order to select reliable training instances from each positive or negative bag. In our new objective function, we balance the weighted training errors and a l1-l2 mixed-norm regularization term which adaptively selects reliable shots as training instances from different videos to have them as diverse as possible. We also develop an alternating optimization approach that can efficiently solve our proposed objective function. Extensive experiments on the challenging real-world Multimedia Event Detection (MED) datasets MEDTest-14, MEDTest-13 and CCV clearly demonstrate the effectiveness of our proposed MIL approach for complex event detection.

****

Deep Direct Regression for Multi-Oriented Scene Text Detection

Wenhao He, Xu-Yao Zhang, Fei Yin, Cheng-Lin Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 745-753

In this paper, we first provide a new perspective to divide existing high performance object detection methods into direct and indirect regressions. Direct regression performs boundary regression by predicting the offsets from a given point, while indirect regression predicts the offsets from some bounding box proposals. In the context of multi-oriented scene text detection, we analyze the drawbacks of indirect regression, which covers the state-of-the-art detection structures Faster-RCNN and SSD as instances, and point out the potential superiority of direct regression. To verify this point of view, we propose a deep direct regression based method for multi-oriented scene text detection. Our detection framework is simple and effective with a fully convolutional network and one-step post processing. The fully convolutional network is optimized in an end-to-end way and has bi-task outputs where one is pixel-wise classification between text and non-text, and the other is direct regression to determine the vertex coordinates of quadrilateral text boundaries. The proposed method is particularly beneficial to localize incidental scene texts. On the ICDAR2015 Incidental Scene Text benchmark, our method achieves the F-measure of 81%, which is a new state-of-the-art and significantly outperforms previous approaches. On other standard datasets with focused scene texts, our method also reaches the state-of-the-art performance.

****

Open Set Domain Adaptation

Pau Panareda Busto, Juergen Gall; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 754-763

When the training and the test data belong to different domains, the accuracy of an object classifier is significantly reduced. Therefore, several algorithms have been proposed in the last years to diminish the so called domain shift between datasets. However, all available evaluation protocols for domain adaptation de

scribe a closed set recognition task, where both domains, namely source and target, contain exactly the same object classes. In this work, we also explore the field of domain adaptation in open sets, which is a more realistic scenario where only a few categories of interest are shared between source and target data. Therefore, we propose a method that fits in both closed and open set scenarios. The approach learns a mapping from the source to the target domain by jointly solving an assignment problem that labels those target instances that potentially belong to the categories of interest present in the source dataset. A thorough evaluation shows that our approach outperforms the state-of-the-art.
****

Deformable Convolutional Networks

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764-773

Convolutional neural networks (CNNs) are inherently limited to model geometric transformations due to the fixed geometric structures in its building modules. In this work, we introduce two new modules to enhance the transformation modeling capacity of CNNs, namely, deformable convolution and deformable RoI pooling. Both are based on the idea of augmenting the spatial sampling locations in the modules with additional offsets and learning the offsets from target tasks, without additional supervision. The new modules can readily replace their plain counterparts in existing CNNs and can be easily trained end-to-end by standard back-propagation, giving rise to deformable convolutional networks. Extensive experiments validate the effectiveness of our approach on sophisticated vision tasks of object detection and semantic segmentation. The code would be released.
****

Ensemble Diffusion for Retrieval

Song Bai, Zhichao Zhou, Jingdong Wang, Xiang Bai, Longin Jan Latecki, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 774-783

As a postprocessing procedure, diffusion process has demonstrated its ability of substantially improving the performance of various visual retrieval systems. Whereas, great efforts are also devoted to similarity (or metric) fusion, seeing that only one individual type of similarity cannot fully reveal the intrinsic relationship between objects. This stimulates a great research interest of considering similarity fusion in the framework of diffusion process (i.e., fusion with diffusion) for robust retrieval. In this paper, we firstly revisit representative methods about fusion with diffusion, and provide new insights which are ignored by previous researchers. Then, observing that existing algorithms are susceptible to noisy similarities, the proposed Regularized Ensemble Diffusion (RED) is bundled with an automatic weight learning paradigm, so that the negative impacts of noisy similarities are suppressed. At last, we integrate several recently-proposed similarities with the proposed framework. The experimental results suggest that we can achieve new state-of-the-art performances on various retrieval tasks, including 3D shape retrieval on ModelNet dataset, and image retrieval on Holidays and Ukbench dataset.
****

FoveaNet: Perspective-Aware Urban Scene Parsing

Xin Li, Zequn Jie, Wei Wang, Changsong Liu, Jimei Yang, Xiaohui Shen, Zhe Lin, Qiang Chen, Shuicheng Yan, Jiashi Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 784-792

Parsing urban scene images is critical for self-driving. Most of current solutions employ generic image parsing models that treat all scales and locations in the images equally and do not consider the geometry property of car-captured urban scene images. Thus, they suffer from heterogeneous object scales caused by perspective projection of cameras on actual scenes and inevitably encounter parsing failures on distant objects as well as other boundary and recognition errors. In this work, we propose a new FoveaNet model to fully exploit the perspective geometry of scene images and address the common failures of generic parsing models. FoveaNet estimates the perspective geometry of a scene image through a convolut ional network which integrates supportive evidence from contextual objects within the image. Based on the perspective geometry information, FoveaNet "undoes" the camera perspective projection--analyzing regions in the space of the actual scene, and thus provides much more reliable parsing results. Furthermore, to effectively address the recognition errors, FoveaNet introduces a new dense CRF model that takes the perspective geometry as a prior potential. We evaluate FoveaNet on two urban scene parsing datasets, Cityspaces and CamVid, which demonstrates that FoveaNet can outperform all the well-established baselines and provide new state-of-the-art performance.

****

Beyond Planar Symmetry: Modeling Human Perception of Reflection and Rotation Symmetries in the Wild

Christopher Funk, Yanxi Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 793-803

Humans take advantage of real world symmetries for various tasks, yet capturing their superb symmetry perception mechanism with a computational model remains elusive. Motivated by a new study demonstrating the extremely high inter-person accuracy of human perceived symmetries in the wild, we have constructed the first deep-learning neural network for reflection and rotation symmetry detection (Sym-NET), trained on photos from MS-COCO (Microsoft-Common Object in COntext) dataset with nearly 11K consistent symmetry-labels from more than 400 human observers. We employ novel methods to convert discrete human labels into symmetry heatmaps, capture symmetry densely in an image and quantitatively evaluate Sym-NET against multiple existing computer vision algorithms. On CVPR 2013 symmetry competition testsets and unseen MS-COCO photos, Sym-NET significantly outperforms all other competitors. Beyond mathematically well-defined symmetries on a plane, Sym-NET demonstrates abilities to identify viewpoint-varied 3D symmetries, partially occluded symmetrical objects, and symmetries at a semantic level.

****

Learning to Reason: End-To-End Module Networks for Visual Question Answering

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 804-813

Natural language questions are inherently compositional, and many are most easily answered by reasoning about their decomposition into modular sub-problems. For example, to answer "is there an equal number of balls and boxes?" we can look for balls, look for boxes, count them, and compare the results. The recently proposed Neural Module Network (NMN) architecture implements this approach to question answering by parsing questions into linguistic substructures and assembling question-specific deep networks from smaller modules that each solve one subtask. However, existing NMN implementations rely on brittle off-the-shelf parsers, and are restricted to the module configurations proposed by these parsers rather than learning them from data. In this paper, we propose End-to-End Module Networks (N2NMNs), which learn to reason by directly predicting instance-specific network layouts without the aid of a parser. Our model learns to generate network structures (by imitating expert demonstrations) while simultaneously learning network parameters (using the downstream task loss). Experimental results on the new CLEVR dataset targeted at compositional question answering show that N2NMNs achieve an error reduction of nearly 50% relative to state-of-the-art attentional approaches, while discovering interpretable network architectures specialized for each question.

****

Hard-Aware Deeply Cascaded Embedding

Yuhui Yuan, Kuiyuan Yang, Chao Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 814-823

Riding on the waves of deep neural networks, deep metric learning has achieved promising results in various tasks by using triplet network or Siamese network. Though the basic goal of making images from the same category closer than the ones from different categories is intuitive, it is hard to optimize the objective directly due to the quadratic or cubic sample size. Hard example mining is widely used to solve the problem, which spends the expensive computation on a subset of samples that are considered hard. However, hard is defined relative to a specific model. Then complex models will treat most samples as easy ones and vice versa for simple models, both of which are not good for training. It is difficult to define a model with the just right complexity and choose hard examples adequately as different samples are of diverse hard levels. This motivates us to propose the novel framework named Hard-Aware Deeply Cascaded Embedding(HDC) to ensemble a set of models with different complexities in cascaded manner to mine hard examples at multiple levels. A sample is judged by a series of models with increasing complexities and only updates models that consider the sample as a hard case. The HDC is evaluated on CARS196, CUB-200-2011, Stanford Online Products, VehicleID and DeepFashion datasets, and outperforms state-of-the-art methods by a large margin.

****

Query-Guided Regression Network With Context Policy for Phrase Grounding

Kan Chen, Rama Kovvuri, Ram Nevatia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 824-832

Given a textual description of an image, phrase grounding localizes objects in the image referred by query phrases in the description. State-of-the-art methods address the problem by ranking a set of proposals based on the relevance to each query, which are limited by the performance of independent proposal generation systems and ignore useful cues from context in the description. In this paper, we adopt a spatial regression method to break the performance limit, and introduce reinforcement learning techniques to further leverage semantic context information. We propose a novel Query-guided Regression network with Context policy (QRC Net) which jointly learns a Proposal Generation Network (PGN), a Query-guided Regression Network (QRN) and a Context Policy Network (CPN). Experiments show QRC Net provides a significant improvement in accuracy on two popular datasets: Flickr30K Entities and Referit Game, with 14.25% and 17.14% increase over the state-of-the-arts respectively.

****

SUBIC: A Supervised, Structured Binary Code for Image Search

Himalaya Jain, Joaquin Zepeda, Patrick Perez, Remi Gribonval; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 833-842

For large-scale visual search, highly compressed yet meaningful representations of images are essential. Structured vector quantizers based on product quantization and its variants are usually employed to achieve such compression while minimizing the loss of accuracy. Yet, unlike binary hashing schemes, these unsupervised methods have not yet benefited from the supervision, end-to-end learning and novel architectures ushered in by the deep learning revolution. We hence propose herein a novel method to make deep convolutional neural networks produce supervised, compact, structured binary codes for visual search. Our method makes use of a novel block-softmax non-linearity and of batch-based entropy losses that together induce structure in the learned encodings. We show that our method outperforms state-of-the-art compact representations based on deep hashing or structured quantization in single and cross-domain category retrieval, instance retrieval and classification. We make our code and models publicly available online.

****

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

Chen Sun, Abhinav Shrivastava, Saurabh Singh, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 843-852

The success of deep learning in vision can be attributed to: (a) models with high capacity; (b) increased computational power; and (c) availability of large-scale labeled data. Since 2012, there have been significant advances in representation capabilities of the models and computational capabilities of GPUs. But the size of the biggest dataset has surprisingly remained constant. What will happen if we increase the dataset size by 10x or 100x? This paper takes a step towards clearing the clouds of mystery surrounding the relationship between `enormous data' and visual deep learning. By exploiting the JFT-300M dataset which has more than 375M noisy labels for 300M images, we investigate how the performance of cu rrent vision tasks would change if this data was used for representation learning. Our paper delivers some surprising (and some expected) findings. First, we find that the performance on vision tasks increases logarithmically based on volume of training data size. Second, we show that representation learning (or pre-training) still holds a lot of promise. One can improve performance on many vision tasks by just training a better base model. Finally, as expected, we present new state-of-the-art results for different vision tasks including image classification, object detection, semantic segmentation and human pose estimation. Our sincere hope is that this inspires vision community to not undervalue the data and develop collective efforts in building larger datasets.

****

A Generative Model of People in Clothing

Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 853-862

We present the first image-based generative model of people in clothing for the full body. We sidestep the commonly used complex graphics rendering pipeline and the need for high-quality 3D scans of dressed people. Instead, we learn generative models from a large image database. The main challenge is to cope with the high variance in human pose, shape and appearance. For this reason, pure image-based approaches have not been considered so far. We show that this challenge can be overcome by splitting the generating process in two parts. First, we learn to generate a semantic segmentation of the body and clothing. Second, we learn a conditional model on the resulting segments that creates realistic images. The full model is differentiable and can be conditioned on pose, shape or color. The result are samples of people in different clothing items and styles. The proposed model can generate entirely new people with realistic clothing. In several experiments we present encouraging results that suggest an entirely data-driven approach to people generation is possible.

****

Escape From Cells: Deep Kd-Networks for the Recognition of 3D Point Cloud Models

Roman Klokov, Victor Lempitsky; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 863-872

We present a new deep learning architecture (called Kd-network) that is designed for 3D model recognition tasks and works with unstructured point clouds. The new architecture performs multiplicative transformations and shares parameters of these transformations according to the subdivisions of the point clouds imposed onto them by kd-trees. Unlike the currently dominant convolutional architectures that usually require rasterization on uniform two-dimensional or three-dimensional grids, Kd-networks do not rely on such grids in any way and therefore avoid poor scaling behavior. In a series of experiments with popular shape recognition benchmarks, Kd-networks demonstrate competitive performance in a number of shape recognition tasks such as shape classification, shape retrieval and shape part segmentation.

****

Improved Image Captioning via Policy Gradient Optimization of SPIDEr

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, Kevin Murphy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 873-881

Current image captioning methods are usually trained via maximum likelihood estimation. However, the log-likelihood score of a caption does not correlate well with human assessments of quality. Standard syntactic evaluation metrics, such as BLEU, METEOR and ROUGE, are also not well correlated. The newer SPICE and CIDEr metrics are better correlated, but have traditionally been hard to optimize for. In this paper, we show how to use a policy gradient (PG) method to directly optimize a linear combination of SPICE and CIDEr (a combination we call SPIDEr): the SPICE score ensures our captions are semantically faithful to the image, while CIDEr score ensures our captions are syntactically fluent. The PG method we propose improves on the prior MIXER approach, by using Monte Carlo rollouts instead of mixing MLE training with PG. We show empirically that our algorithm leads to easier optimization and improved results compared to MIXER. Finally, we show that using our PG method we can optimize any of the metrics, including the propos ed SPIDEr metric which results in image captions that are strongly preferred by human raters compared to captions generated by the same model but trained to optimize MLE or the COCO metrics.

****

Rolling Shutter Correction in Manhattan World

Pulak Purkait, Christopher Zach, Ales Leonardis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 882-890

A vast majority of consumer cameras operate the rolling shutter mechanism, which often produces distorted images due to inter-row delay while capturing an image. Recent methods for monocular rolling shutter compensation utilize blur kernel, straightness of line segments, as well as angle and length preservation. However, they do not incorporate scene geometry explicitly for rolling shutter correction, therefore, information about the 3D scene geometry is often distorted by the correction process. In this paper we propose a novel method which leverages geometric properties of the scene---in particular vanishing directions---to estimate the camera motion during rolling shutter exposure from a single distorted image. The proposed method jointly estimates the orthogonal vanishing directions and the rolling shutter camera motion. We performed extensive experiments on synthetic and real datasets which demonstrate the benefits of our approach both in terms of qualitative and quantitative results (in terms of a geometric structure fitting) as well as with respect to computation time.

****

Local-To-Global Point Cloud Registration Using a Dictionary of Viewpoint Descriptors

David Avidar, David Malah, Meir Barzohar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 891-899

Local-to global point cloud registration is a challenging task due to the substantial differences between these two types of data, and the different techniques used to acquire them. Global clouds cover large-scale environments and are usually acquired aerially, e.g., 3D modeling of a city using Airborne Laser Scanning (ALS). In contrast, local clouds are often acquired from ground level and at a much smaller range, for example, using Terrestrial Laser Scanning (TLS). The differences are often manifested in point density distribution, occlusions nature, and measurement noise. As a result of these differences, existing point cloud registration approaches, such as keypoint-based registration, tend to fail. We improve upon a different approach, recently proposed, based on converting the global cloud into a viewpoint-based cloud dictionary. We propose a local-to-global registration method where we replace the dictionary clouds with viewpoint descriptors, consisting of panoramic range-images. We then use an efficient dictionary search in the Discrete Fourier Transform (DFT) domain, using phase correlation, to rapidly find plausible transformations from the local to the global reference frame. We demonstrate our method's significant advantages over the previous cloud dictionary approach, in terms of computational efficiency and memory requirements. In addition, We show its superior registration performance in comparison to a state-of-the-art, keypoint-based method (FPFH). For the evaluation, we use a challenging dataset of TLS local clouds and an ALS large-scale global cloud, in an urban environment.

****

3D-PRNN: Generating Shape Primitives With Recurrent Neural Networks

Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, Derek Hoiem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 900-909

The success of various applications including robotics, digital content creation, and visualization demand a structured and abstract representation of the 3D world from limited sensor data. Inspired by the nature of human perception of 3D shapes as a collection of simple parts, we explore such an abstract shape representation based on primitives. Given a single depth image of an object, we present 3D-PRNN, a generative recurrent neural network that synthesizes multiple plausible shapes composed of a set of primitives. Our generative model encodes symmetry characteristics of common man-made objects, preserves long-range structural coherence, and describes objects of varying complexity with a compact representati on. We also propose a method based on Gaussian Fields to generate a large scale dataset of primitive-based shape representations to train our network. We evaluate our approach on a wide range of examples and show that it outperforms nearest-neighbor based shape retrieval methods and is on-par with voxel-based generative models while using a significantly reduced parameter space.

****

BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera

Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, Yebin Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 910-919

We propose BodyFusion, a novel real-time geometry fusion method that can track and reconstruct non-rigid surface motion of a human performance using a single consumer-grade depth camera. To reduce the ambiguities of the non-rigid deformation parameterization on the surface graph nodes, we take advantage of the internal articulated motion prior for human performance and contribute a skeleton-embedded surface fusion (SSF) method. The key feature of our method is that it jointly solves for both the skeleton and graph-node deformations based on information of the attachments between the skeleton and the graph nodes. The attachments are also updated frame by frame based on the fused surface geometry and the computed deformations. Overall, our method enables increasingly denoised, detailed, and complete surface reconstruction as well as the updating of the skeleton and attachments as the temporal depth frames are fused. Experimental results show that our method exhibits substantially improved nonrigid motion fusion performance and tracking robustness compared with previous state-of-the-art fusion methods. We also contribute a dataset for the quantitative evaluation of fusion-based dynamic scene reconstruction algorithms using a single depth camera.

****

Quasiconvex Plane Sweep for Triangulation With Outliers

Qianggong Zhang, Tat-Jun Chin, David Suter; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 920-928

Triangulation is a fundamental task in 3D computer vision. Unsurprisingly, it is a well-investigated problem with many mature algorithms. However, algorithms for robust triangulation, which are necessary to produce correct results in the presence of egregiously incorrect measurements (i.e., outliers), have received much less attention. The default approach to deal with outliers in triangulation is by random sampling. The randomized heuristic is not only suboptimal, it could, in fact, be computationally inefficient on large-scale datasets. In this paper, we propose a novel locally optimal algorithm for robust triangulation. A key feature of our method is to efficiently derive the local update step by plane sweeping a set of quasiconvex functions. Underpinning our method is a new theory behind quasiconvex plane sweep, which has not been examined previously in computational geometry. Relative to the random sampling heuristic, our algorithm not only guarantees deterministic convergence to a local minimum, it typically achieves higher quality solutions in similar runtimes.

****

"Maximizing Rigidity" Revisited: A Convex Programming Approach for Generic 3D Shape Reconstruction From Multiple Perspective Views

Pan Ji, Hongdong Li, Yuchao Dai, Ian Reid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 929-937

Rigid structure-from-motion (RSfM) and non-rigid structure-from-motion (NRSfM) have long been treated in the literature as separate (different) problems. Inspired by a previous work which solved directly for 3D scene structure by factoring the relative camera poses out, we revisit the principle of "maximizing rigidity" in structure-from-motion literature, and develop a unified theory which is applicable to both rigid and non-rigid structure reconstruction in a rigidity-agnostic way. We formulate these problems as a convex semi-definite program, imposing constraints that seek to apply the principle of minimizing non-rigidity. Our results demonstrate the efficacy of the approach, with state-of-the-art accuracy on various 3D reconstruction problems.

****

Surface Registration via Foliation

Xiaopeng Zheng, Chengfeng Wen, Na Lei, Ming Ma, Xianfeng Gu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 938-947

This work introduces a novel surface registration method based on foliation. A foliation decomposes the surface into a family of closed loops, such that the decomposition has local tensor product structure. By projecting each loop to a point, the surface is collapsed into a graph. Two homeomorphic surfaces with consistent foliations can be registered by first matching their foliation graphs, then matching the corresponding leaves. This foliation based method is capable of handling surfaces with complicated topologies and large non-isometric deformations, rigorous with solid theoretic foundation, easy to implement, robust to compute. The result mapping is diffeomorphic. Our experimental results show the efficiency and efficacy of the proposed method.

****

Rolling-Shutter-Aware Differential SfM and Image Rectification

Bingbing Zhuang, Loong-Fah Cheong, Gim Hee Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 948-956

In this paper, we develop a modified differential Structure from Motion (SfM) algorithm that can estimate relative pose from two frames despite of Rolling Shutter (RS) artifacts. In particular, we show that under constant velocity assumption, the errors induced by the rolling shutter effect can be easily rectified by a linear scaling operation on each optical flow. We further propose a 9-point algorithm to recover the relative pose of a rolling shutter camera that undergoes constant acceleration motion. We demonstrate that the dense depth maps recovered from the relative pose of the RS camera can be used in a RS-aware warping for image rectification to recover high-quality Global Shutter (GS) images. Experiments on both synthetic and real RS images show that our RS-aware differential SfM algorithm produces more accurate results on relative pose estimation and 3D reconstruction from images distorted by RS effect compared to standard SfM algorithms that assume a GS camera model. We also demonstrate that our RS-aware warping for image rectification method outperforms state-of-the-art commercial software products, i.e. Adobe After Effects and Apple Imovie, at removing RS artifacts.

****

Corner-Based Geometric Calibration of Multi-Focus Plenoptic Cameras

Sotiris Nousias, Francois Chadebecq, Jonas Pichat, Pearse Keane, Sebastien Ourselin, Christos Bergeles; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 957-965

We propose a method for geometric calibration of multi-focus plenoptic cameras using raw images. Multi-focus plenoptic cameras feature several types of micro-lenses spatially aligned in front of the camera sensor to generate micro-images at different magnifications. This multi-lens arrangement provides computational-photography benefits but complicates calibration. Our methodology achieves the detection of the type of micro-lenses, the retrieval of their spatial arrangement, and the estimation of intrinsic and extrinsic camera parameters therefore fully characterising this specialised camera class. Motivated from classic pinhole camera calibration, the presented algorithm operates based on a checker-board's corners, retrieved by a custom micro-image corner detector. This approach enables the introduction of a re-projection error that is used in a minimisation framework. Our algorithm compares favourably to the state-of-the-art, as demonstrated by controlled and free-hand experiments, making it a first step towards accurate 3D reconstruction and Structure-from-Motion.

****

Focal Track: Depth and Accommodation With Oscillating Lens Deformation

Qi Guo, Emma Alexander, Todd Zickler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 966-974

The focal track sensor is a monocular and computationally efficient depth sensor that is based on defocus controlled by a liquid membrane lens. It synchronizes small lens oscillations with a photosensor to produce real-time depth maps by means of differential defocus, and it couples these oscillations with bigger lens deformations that adapt the defocus working range to track objects over large axial distances. To create the focal track sensor, we derive a texture-invariant family of equations that relate image derivatives to scene depth when a lens changes its focal length differentially. Based on these equations, we design a feed-forward sequence of computations that: robustly incorporates image derivatives at multiple scales; produces confidence maps along with depth; and can be trained end-to-end to mitigate against noise, aberrations, and other non-idealities. Our prototype with 1-inch optics produces depth and confidence maps at 100 frames per second over an axial range of more than 75cm.

****

Reconfiguring the Imaging Pipeline for Computer Vision

Mark Buckler, Suren Jayasuriya, Adrian Sampson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 975-984

Advancements in deep learning have ignited an explosion of research on efficient hardware for embedded computer vision. Hardware vision acceleration, however, does not address the cost of capturing and processing the image data that feeds these algorithms. We examine the role of the image signal processing (ISP) pipeline in computer vision to identify opportunities to reduce computation and save energy. The key insight is that imaging pipelines should be designed to be configurable: to switch between a traditional photography mode and a low-power vision mode that produces lower-quality image data suitable only for computer vision. We use eight computer vision algorithms and a reversible pipeline simulation tool to study the imaging system's impact on vision performance. For both CNN-based and classical vision algorithms, we observe that only two ISP stages, demosaicing and gamma compression, are critical for task performance. We propose a new image sensor design that can compensate for skipping these stages. The sensor design features an adjustable resolution and tunable analog-to-digital converters (ADCs). Our proposed imaging system's vision mode disables the ISP entirely and configures the sensor to produce subsampled, lower-precision image data. This vision mode can save 75% of the average energy of a baseline photography mode while having only a small impact on vision task accuracy.

****

Catadioptric HyperSpectral Light Field Imaging

Yujia Xue, Kang Zhu, Qiang Fu, Xilin Chen, Jingyi Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 985-993

The complete plenoptic function records radiance of rays from every location, at every angle, for every wavelength and at every time. The signal is multi-dimensional and has long relied on multi-modal sensing such as hybrid light field camera arrays. In this paper, we present a single camera hyperspectral light field imaging solution that we call Snapshot Plenoptic Imager (SPI). SPI uses spectral coded catadioptric mirror arrays for simultaneously acquiring the spatial, angular and spectral dimensions. We further apply a learning-based approach to improve the spectral resolution from very few measurements. Specifically, we demonstrate and then employ a new spectral sparsity prior that allows the hyperspectral profiles to be sparsely represented under a pre-trained dictionary. Comprehensive experiments on synthetic and real data show that our technique is effective, reliable, and accurate. In particular, we are able to produce the first wide FoV multi-spectral light field database.

****

Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification

Hong-Xing Yu, Ancong Wu, Wei-Shi Zheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 994-1002

While metric learning is important for Person re-identification (RE-ID), a significant problem in visual surveillance for cross-view pedestrian matching, existing metric models for RE-ID are mostly based on supervised learning that requires quantities of labeled samples in all pairs of camera views for training. However, this limits their scalabilities to realistic applications, in which a large amount of data over multiple disjoint camera views is available but not labelled. To overcome the problem, we propose an unsupervised asymmetric metric learning model for unsupervised RE-ID. Our model aims to learn an asymmetric metric, i.e.

, specific projection for each view, effectively based on clustering on cross-view person images. Our model finds a shared space where view-specific bias is alleviated and thus better matching performance can be achieved. Extensive experiments have been conducted on a baseline and five large-scale RE-ID datasets to demonstrate the effectiveness of the proposed model. Through the comparison, we show that our unsupervised asymmetric metric model works much more suitable for unsupervised RE-ID as compared to classical unsupervised metric learning models. We also compare existing unsupervised RE-ID methods, and our model outperforms them with notable margins, and especially we report the performance on large-scale unlabelled RE-ID dataset, which is unfortunately less concerned in literatures.
****

Real Time Eye Gaze Tracking With 3D Deformable Eye-Face Model
Kang Wang, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1003-1011

3D model-based gaze estimation methods are widely explored because of their good accuracy and ability to handle free head movement. Traditional methods with complex hardware systems (Eg. infrared lights, 3D sensors, etc.) are restricted to controlled environments, which significantly limit their practical utilities. In this paper, we propose a 3D model-based gaze estimation method with a single web-camera, which enables instant and portable eye gaze tracking. The key idea is to leverage on the proposed 3D eye-face model, from which we can estimate 3D eye gaze from observed 2D facial landmarks. The proposed system includes a 3D deformable eye-face model that is learned offline from multiple training subjects. Given the deformable model, individual 3D eye-face models and personal eye parameters can be recovered through the unified calibration algorithm. Experimental results show that the proposed method outperforms state-of-the-art methods while allowing convenient system setup and free head movement. A real time eye tracking system running at 30 FPS also validates the effectiveness and efficiency of the proposed method.
****

Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks
Inwoong Lee, Doyoung Kim, Seoungyoon Kang, Sanghoon Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1012-1020

This paper addresses the problems of feature representation of skeleton joints and the modeling of temporal dynamics to recognize human actions. Traditional methods generally use relative coordinate systems dependent on some joints, and model only the long-term dependency, while excluding short-term and medium term dependencies. Instead of taking raw skeletons as the input, we transform the skeletons into another coordinate system to obtain the robustness to scale, rotation and translation, and then extract salient motion features from them. Considering that Long Short-term Memory (LSTM) networks with various time-step sizes can model various attributes well, we propose novel ensemble Temporal Sliding LSTM (TS-LSTM) networks for skeleton-based action recognition. The proposed network is composed of multiple parts containing short-term, medium-term and long-term TS-LSTM networks, respectively. In our network, we utilize an average ensemble among multiple parts as a final feature to capture various temporal dependencies. We evaluate the proposed networks and the additional other architectures to verify the effectiveness of the proposed networks, and also compare them with several other methods on five challenging datasets. The experimental results demonstrate that our network models achieve the state-of-the-art performance through various temporal features. Additionally, we analyze a relation between the recognized actions and the multi-term TS-LSTM features by visualizing the softmax features of multiple parts.
****

How Far Are We From Solving the 2D & 3D Face Alignment Problem? (And a Dataset of 230,000 3D Facial Landmarks)
Adrian Bulat, Georgios Tzimiropoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1021-1030

This paper investigates how far a very deep neural network is from attaining clo se to saturating performance on existing 2D and 3D face alignment datasets. To this end, we make the following 5 contributions: (a) we construct, for the first time, a very strong baseline by combining a state-of-the-art architecture for landmark localization with a state-of-the-art residual block, train it on a very large yet synthetically expanded 2D facial landmark dataset and finally evaluate it on all other 2D facial landmark datasets. (b) We create a guided by 2D landmarks network which converts 2D landmark annotations to 3D and unifies all existing datasets, leading to the creation of LS3D-W, the largest and most challenging 3D facial landmark dataset to date 230,000 images. (c) Following that, we train a neural network for 3D face alignment and evaluate it on the newly introduced LS3D-W. (d) We further look into the effect of all "traditional" factors affecting face alignment performance like large pose, initialization and resolution, and introduce a "new" one, namely the size of the network. (e) We show that both 2D and 3D face alignment networks achieve performance of remarkable accuracy which is probably close to saturating the datasets used. Training and testing code as well as the dataset can be downloaded from https://www.adrianbulat.com/face-alignment/

****

Large Pose 3D Face Reconstruction From a Single Image via Direct Volumetric CNN Regression

Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, Georgios Tzimiropoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1031-1039

3D face reconstruction is a fundamental Computer Vision problem of extraordinary difficulty. Current systems often assume the availability of multiple facial images (sometimes from the same subject) as input, and must address a number of methodological challenges such as establishing dense correspondences across large facial poses, expressions, and non-uniform illumination. In general these methods require complex and inefficient pipelines for model building and fitting. In this work, we propose to address many of these limitations by training a Convolutional Neural Network (CNN) on an appropriate dataset consisting of 2D images and 3D facial models or scans. Our CNN works with just a single 2D facial image, does not require accurate alignment nor establishes dense correspondence between images, works for arbitrary facial poses and expressions, and can be used to reconstruct the whole 3D facial geometry (including the non-visible parts of the face) bypassing the construction (during training) and fitting (during testing) of a 3D Morphable Model. We achieve this via a simple CNN architecture that performs direct regression of a volumetric representation of the 3D facial geometry from a single 2D image. We also demonstrate how the related task of facial landmark localization can be incorporated into the proposed framework and help improve reconstruction quality, especially for the cases of large poses and facial expressions.

****

RankIQA: Learning From Rankings for No-Reference Image Quality Assessment

Xialei Liu, Joost van de Weijer, Andrew D. Bagdanov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1040-1049

We propose a no-reference image quality assessment (NR-IQA) approach that learns from rankings (RankIQA). To address the problem of limited IQA dataset size, we train a Siamese Network to rank images in terms of image quality by using synthetically generated distortions for which relative image quality is known. These ranked image sets can be automatically generated without laborious human labeling. We then use fine-tuning to transfer the knowledge represented in the trained Siamese Network to a traditional CNN that estimates absolute image quality from single images. We demonstrate how our approach can be made significantly more efficient than traditional Siamese Networks by forward propagating a batch of images through a single network and backpropagating gradients derived from all pairs of images in the batch. Experiments on the TID2013 benchmark show that we improve the state-of-the-art by over 5%. Furthermore, on the LIVE benchmark we show that our approach is superior to existing NR-IQA techniques and that we even outperform the state-of-the-art in full-reference IQA (FR-IQA) methods without havin g to resort to high-quality reference images to infer IQA.

****

Look, Perceive and Segment: Finding the Salient Objects in Images via Two-Stream Fixation-Semantic CNNs

Xiaowu Chen, Anlin Zheng, Jia Li, Feng Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1050-1058

Recently, CNN-based models have achieved remarkable success in image-based salient object detection (SOD). In these models, a key issue is to find a proper network architecture that best fits for the task of SOD. Toward this end, this paper proposes two-stream fixation-semantic CNNs, whose architecture is inspired by the fact that salient objects in complex images can be unambiguously annotated by selecting the pre-segmented semantic objects that receive the highest fixation density in eye-tracking experiments. In the two-stream CNNs, a fixation stream is pre-trained on eye-tracking data whose architecture well fits for the task of fixation prediction, and a semantic stream is pre-trained on images with semantic tags that has a proper architecture for semantic perception. By fusing these two streams into an inception-segmentation module and jointly fine-tuning them on images with manually annotated salient objects, the proposed networks show impressive performance in segmenting salient objects. Experimental results show that our approach outperforms 10 state-of-the-art models (5 deep, 5 non-deep) on 4 datasets.

****

Delving Into Salient Object Subitizing and Detection

Shengfeng He, Jianbo Jiao, Xiaodan Zhang, Guoqiang Han, Rynson W.H. Lau; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1059-1067

Subitizing (i.e., instant judgement on the number) and detection of salient objects are human inborn abilities. These two tasks influence each other in the human visual system. In this paper, we delve into the complementarity of these two tasks. We propose a multi-task deep neural network with weight prediction for salient object detection, where the parameters of an adaptive weight layer are dynamically determined by an auxiliary subitizing network. The numerical representation of salient objects is therefore embedded into the spatial representation. The proposed joint network can be trained end-to-end using back-propagation. Experiments show that the proposed multi-task network outperforms existing multi-task architectures, and the auxiliary subitizing network provides strong guidance to salient object detection by reducing false positives and producing coherent saliency maps. Moreover, the proposed method is an unconstrained method able to handle images with/without salient objects. Finally, we show state-of-theart performance on different salient object datasets.

****

Learning Discriminative Data Fitting Functions for Blind Image Deblurring

Jinshan Pan, Jiangxin Dong, Yu-Wing Tai, Zhixun Su, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1068-1076

Solving blind image deblurring usually requires defining a data fitting function and image priors. While existing algorithms mainly focus on developing image priors for blur kernel estimation and non-blind deconvolution, only a few methods consider the effect of data fitting functions. In contrast to the state-of-the-art methods that use a single or a fixed data fitting term, we propose a data-driven approach to learn effective data fitting functions from a large set of motion blurred images with associated ground truth blur kernels. The learned data fitting function facilitates estimating accurate blur kernels for generic images and domain-specific problems with corresponding image priors. In addition, we extend the learning approach for data fitting function to latent image restoration and non-uniform deblurring. Extensive experiments on challenging motion blurred images demonstrate the proposed algorithm performs favorably against the state-of-the-art methods.

****

Video Deblurring via Semantic Segmentation and Pixel-Wise Non-Linear Kernel

Wenqi Ren, Jinshan Pan, Xiaochun Cao, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1077-1085

Video deblurring is a challenging problem as the blur is complex and usually caused by the combination of camera shakes, object motions, and depth variations. Optical flow can be used for kernel estimation since it predicts motion trajectories. However, the estimates are often inaccurate in complex scenes at object boundaries, which are crucial in kernel estimation. In this paper, we exploit semantic segmentation in each blurry frame to understand the scene contents and use different motion models for image regions to guide optical flow estimation. While existing pixel-wise blur models assume that the blur kernel is the same as optical flow during the exposure time, this assumption does not hold when the motion blur trajectory at a pixel is different from the estimated linear optical flow. We analyze the relationship between motion blur trajectory and optical flow, and present a novel pixel-wise non-linear kernel model to account for motion blur. The proposed blur model is based on the non-linear optical flow, which describes complex motion blur more effectively. Extensive experiments on challenging blurry videos demonstrate the proposed algorithm performs favorably against the state-of-the-art methods.

****

On-Demand Learning for Deep Image Restoration

Ruohan Gao, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1086-1095

While machine learning approaches to image restoration offer great promise, current methods risk training models fixated on performing well only for image corruption of a particular level of difficulty--such as a certain level of noise or blur. First, we examine the weakness of conventional "fixated" models and demonstrate that training general models to handle arbitrary levels of corruption is indeed non-trivial. Then, we propose an on-demand learning algorithm for training image restoration models with deep convolutional neural networks. The main idea is to exploit a feedback mechanism to self-generate training instances where they are needed most, thereby learning models that can generalize across difficulty levels. On four restoration tasks--image inpainting, pixel interpolation, image deblurring, and image denoising--and three diverse datasets, our approach consistently outperforms both the status quo training procedure and curriculum learning alternatives.

****

Multi-Channel Weighted Nuclear Norm Minimization for Real Color Image Denoising

Jun Xu, Lei Zhang, David Zhang, Xiangchu Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1096-1104

Most of the existing denoising algorithms are developed for grayscale images. It is not trivial to extend them for color image denoising since the noise statistics in R, G, and B channels can be very different for real noisy images. In this paper, we propose a multi-channel (MC) optimization model for real color image denoising under the weighted nuclear norm minimization (WNNM) framework. We concatenate the RGB patches to make use of the channel redundancy, and introduce a weight matrix to balance the data fidelity of the three channels in consideration of their different noise statistics. The proposed MC-WNNM model does not have an analytical solution. We reformulate it into a linear equality-constrained problem and solve it via alternating direction method of multipliers. Each alternative updating step has a closed-form solution and the convergence can be guaranteed. Experiments on both synthetic and real noisy image datasets demonstrate the superiority of the proposed MC-WNNM over state-of-the-art denoising methods.

****

Coherent Online Video Style Transfer

Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, Gang Hua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1105-1114

Training a feed-forward network for the fast neural style transfer of images has proven successful, but the naive extension of processing videos frame by frame is prone to producing flickering results. We propose the first end-to-end network for online video style transfer, which generates temporally coherent stylized

video sequences in near real-time. Two key ideas include an efficient network by incorporating short-term coherence, and propagating short-term coherence to long-term, which ensures consistency over a longer period of time. Our network can incorporate different image stylization networks and clearly outperforms the per-frame baseline both qualitatively and quantitatively. Moreover, it can achieve visually comparable coherence to optimization-based video style transfer, but is three orders of magnitude faster.
****

SHaPE: A Novel Graph Theoretic Algorithm for Making Consensus-Based Decisions in Person Re-Identification Systems
Arko Barman, Shishir K. Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1115-1124
Person re-identification is a challenge in video-based surveillance where the goal is to identify the same person in different camera views. In recent years, many algorithms have been proposed that approach this problem by designing suitable feature representations for images of persons or by training appropriate distance metrics that learn to distinguish between images of different persons. Aggregating the results from multiple algorithms for person re-identification is a relatively less-explored area of research. In this paper, we formulate an algorithm that maps the ranking process in a person re-identification algorithm to a problem in graph theory. We then extend this formulation to allow for the use of results from multiple algorithms to make a consensus-based decision for the person re-identification problem. The algorithm is unsupervised and takes into account only the matching scores generated by multiple algorithms for creating a consensus of results. Further, we show how the graph theoretic problem can be solved by a two-step process. First, we obtain a rough estimate of the solution using a greedy algorithm. Then, we extend the construction of the proposed graph so that the problem can be efficiently solved by means of Ant Colony Optimization, a heuristic path-searching algorithm for complex graphs. While we present the algorithm in the context of person re-identification, it can potentially be applied to the general problem of ranking items based on a consensus of multiple sets of scores or metric values.
****

Need for Speed: A Benchmark for Higher Frame Rate Object Tracking
Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1125-1134
In this paper, we propose the first higher frame rate video dataset (called Need for Speed - NfS) and benchmark for visual object tracking. The dataset consists of 100 videos (380K frames) captured with now commonly available higher frame rate (240 FPS) cameras from real world scenarios. All frames are annotated with axis aligned bounding boxes and all sequences are manually labelled with nine visual attributes - such as occlusion, fast motion, background clutter, etc. Our benchmark provides an extensive evaluation of many recent and state-of-the-art trackers on higher frame rate sequences. We ranked each of these trackers according to their tracking accuracy and real-time performance. One of our surprising conclusions is that at higher frame rates, simple trackers such as correlation filters outperform complex methods based on deep networks. This suggests that for practical applications (such as in robotics or embedded vision), one needs to carefully tradeoff bandwidth constraints associated with higher frame rate acquisition, computational costs of real-time analysis, and the required application accuracy. Our dataset and benchmark allows for the first time (to our knowledge) systematic exploration of such issues, and will be made available to allow for further research in this space.
****

Learning Background-Aware Correlation Filters for Visual Tracking
Hamed Kiani Galoogahi, Ashton Fagg, Simon Lucey; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1135-1143
Correlation Filters (CFs) have recently demonstrated excellent performance in terms of rapidly tracking objects under challenging photometric and geometric vari

ations. The strength of the approach comes from its ability to efficiently learn - "on the fly" - how the object is changing over time. A fundamental drawback to CFs, however, is that the background of the target is not modeled over time which can result in suboptimal performance. Recent tracking algorithms have suggested to resolve this drawback by either learning CFs from more discriminative deep features (e.g. DeepSRDCF and CCOT) or learning complex deep trackers (e.g. MDNet and FCNT). While such methods have been shown to work well, they suffer from high complexity: extracting deep features or applying deep tracking frameworks is very computationally expensive. This limits the real-time performance of such methods, even on high-end GPUs. This work proposes a Background-Aware CF based on hand-crafted features (HOG) that can efficiently model how both the foreground and background of the object varies over time. Our approach, like conventional CFs, is extremely computationally efficient- and extensive experiments over multiple tracking benchmarks demonstrate the superior accuracy and real-time performance of our method compared to the state-of-the-art trackers.

****

Robust Object Tracking Based on Temporal and Spatial Deep Networks

Zhu Teng, Junliang Xing, Qiang Wang, Congyan Lang, Songhe Feng, Yi Jin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1144-1153

Recently deep neural networks have been widely employed to deal with the visual tracking problem. In this work, we present a new deep architecture which incorporates the temporal and spatial information to boost the tracking performance. Our deep architecture contains three networks, a Feature Net, a Temporal Net, and a Spatial Net. The Feature Net extracts general feature representations of the target. With these feature representations, the Temporal Net encodes the trajectory of the target and directly learns temporal correspondences to estimate the object state from a global perspective. Based on the learning results of the Temporal Net, the Spatial Net further refines the object tracking state using local spatial object information. Extensive experiments on four of the largest tracking benchmarks, including VOT2014, VOT2016, OTB50, and OTB100, demonstrate competing performance of the proposed tracker over a number of state-of-the-art algorithms.

****

Real-Time Hand Tracking Under Occlusion From an Egocentric RGB-D Sensor

Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1154-1163

We present an approach for real-time, robust, and accurate hand pose estimation from moving egocentric RGB-D cameras in cluttered real environments. Existing methods typically fail for hand-object interactions in cluttered scenes imaged from egocentric viewpoints, common for virtual or augmented reality applications. Our approach uses two subsequently applied Convolutional Neural Networks (CNNs) to localize the hand and regress 3D joint locations. Hand localization is achieved by using a CNN to estimate the 2D position of the hand center in the input, even in the presence of clutter and occlusions. The localized hand position, together with the corresponding input depth value, is used to generate a normalized cropped image that is fed into a second CNN to regress relative 3D hand joint locations in real-time. For added accuracy, robustness, and temporal stability, we refine the pose estimates using a kinematic pose tracking energy. To train the CNNs, we introduce a new photorealistic dataset that uses a merged reality approach to capture and synthesize large amounts of annotated data of natural hand interaction in cluttered scenes. Through quantitative and qualitative evaluation, we show that our method is robust to self-occlusion and occlusions by objects, specifically in moving egocentric perspectives.

****

Predicting Human Activities Using Stochastic Grammar

Siyuan Qi, Siyuan Huang, Ping Wei, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1164-1172

This paper presents a novel method to predict future human activities from parti ally observed RGB-D videos. Human activity prediction is generally difficult due to its non-Markovian property and the rich context between human and environments. We use a stochastic grammar model to capture the compositional structure of events, integrating human actions, objects, and their affordances. We represent the event by a spatial-temporal And-Or graph (ST-AOG). The ST-AOG is composed of a temporal stochastic grammar defined on sub-activities, and spatial graphs representing sub-activities that consist of human actions, objects, and their affordances. Future sub-activities are predicted using the temporal grammar and Earley parsing algorithm. The corresponding action, object, and affordance labels are then inferred accordingly. Extensive experiments are conducted to show the effectiveness of our model on both semantic event parsing and future activity prediction.

****

ProbFlow: Joint Optical Flow and Uncertainty Estimation

Anne S. Wannenwetsch, Margret Keuper, Stefan Roth; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1173-1182

Optical flow estimation remains challenging due to untextured areas, motion boundaries, occlusions, and more. Thus, the estimated flow is not equally reliable across the image. To that end, post-hoc confidence measures have been introduced to assess the per-pixel reliability of the flow. We overcome the artificial separation of optical flow and confidence estimation by introducing a method that jointly predicts optical flow and its underlying uncertainty. Starting from common energy-based formulations, we rely on the corresponding posterior distribution of the flow given the images. We derive a variational inference scheme based on mean field, which incorporates best practices from energy minimization. An uncertainty measure is obtained along the flow at every pixel as the (marginal) entropy of the variational distribution. We demonstrate the flexibility of our probabilistic approach by applying it to two different energies and on two benchmarks. We not only obtain flow results that are competitive with the underlying energy minimization approach, but also a reliable uncertainty measure that significantly outperforms existing post-hoc approaches.

****

Sublabel-Accurate Discretization of Nonconvex Free-Discontinuity Problems

Thomas Mollenhoff, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1183-1191

In this work we show how sublabel-accurate multilabeling approaches can be derived by approximating a classical label-continuous convex relaxation of nonconvex free-discontinuity problems. This insight allows to extend these sublabel-accurate approaches from total variation to general convex and nonconvex regularizations. Furthermore, it leads to a systematic approach to the discretization of continuous convex relaxations. We study the relationship to existing discretizations and to discrete-continuous MRFs. Finally, we apply the proposed approach to obtain a sublabel-accurate and convex solution to the vectorial Mumford-Shah functional and show in several experiments that it leads to more precise solutions using fewer labels.

****

DeepContext: Context-Encoding Neural Pathways for 3D Holistic Scene Understanding

Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, Jianxiong Xiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1192-1201

3D context has been shown to be an extremely important cue for scene understanding, yet very little research has been done on integrating context information with deep models. This paper presents an approach to embed 3D context into the topology of a neural network trained to perform holistic scene understanding. Given a depth image depicting a 3D scene, our network aligns the observed scene with a predefined 3D scene template, and then reasons about the existence and location of each object within the scene template. In doing so, our model recognizes multiple objects in a single forward pass of a 3D convolutional neural network, capturing both global scene and local object information simultaneously. To create

training data for this 3D network, we generate partly hallucinated depth images which are rendered by replacing real objects with a repository of CAD models of the same object category. Extensive experiments demonstrate the effectiveness of our algorithm compared to the state of the art.
****

BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography
Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, Serge Belongie; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1202-1211
Computer vision systems are designed to work well within the context of everyday photography. However, artists often render the world around them in ways that do not resemble photographs. Artwork produced by people is not constrained to mimic the physical world, making it more challenging for machines to recognize. This work is a step toward teaching machines how to categorize images in ways that are valuable to humans. First, we collect a large-scale dataset of contemporary artwork from Behance, a website containing millions of portfolios from professional and commercial artists. We annotate Behance imagery with rich attribute labels for content, emotions, and artistic media. Furthermore, we carry out baseline experiments to show the value of this dataset for artistic style prediction, for improving the generality of existing object classifiers, and for the study of visual domain adaptation. We believe our Behance Artistic Media dataset will be a good starting point for researchers wishing to study artistic imagery and relevant problems. This dataset can be found at https://bam-dataset.org/
****

Adversarial PoseNet: A Structure-Aware Convolutional Network for Human Pose Estimation
Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, Jian Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1212-1221
For human pose estimation in monocular images, joint occlusions and overlapping upon human bodies often result in deviated pose predictions. Under these circumstances, bi- ologically implausible pose predictions may be produced. In contrast, human vision is able to predict poses by exploiting geometric constraints of joint inter-connectivity. To address the problem by incorporating priors about the structure of human bodies, we propose a novel structure-aware convo- lutional network to implicitly take such priors into account during training of the deep network. Explicit learning of such constraints is typically challenging. Instead, we design discriminators to distinguish the real poses from the fake ones (such as biologically implausible ones). If the pose generator (G) generates results that the discriminator fails to distinguish from real ones, the network successfully learns the priors. To better capture the structure dependency of human body joints, the generator G is designed in a stacked multi-task manner to predict poses as well as occlusion heatmaps. Then, the pose and occlusion heatmaps are sent to the discrimina- tors to predict the likelihood of the pose being real. Training of the network follows the strategy of conditional Generative Adversarial Networks (GANs). The effectiveness of the pro- posed network is evaluated on two widely used human pose estimation benchmark datasets. Our approach significantly outperforms the state-of-the-art methods and almost always generates plausible human pose predictions.
****

An Empirical Study of Language CNN for Image Captioning
Jiuxiang Gu, Gang Wang, Jianfei Cai, Tsuhan Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1222-1231
Language models based on recurrent neural networks have dominated recent image caption generation tasks. In this paper, we introduce a Language CNN model which is suitable for statistical language modeling tasks and shows competitive performance in image captioning. In contrast to previous models which predict next word based on one previous word and hidden state, our language CNN is fed with all the previous words and can model the long-range dependencies in history words, which are critical for image captioning. The effectiveness of our approach is validated on two datasets: Flickr30K and MS COCO. Our extensive experimental result

s show that our method outperforms the vanilla recurrent neural network based language models and is competitive with the state-of-the-art methods.

****

Attributes2Classname: A Discriminative Model for Attribute-Based Unsupervised Zero-Shot Learning

Berkan Demirel, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1232-1241

We propose a novel approach for unsupervised zero-shot learning (ZSL) of classes based on their names. Most existing unsupervised ZSL methods aim to learn a model for directly comparing image features and class names. However, this proves to be a difficult task due to dominance of non-visual semantics in underlying vector-space embeddings of class names. To address this issue, we discriminatively learn a word representation such that the similarities between class and combination of attribute names fall in line with the visual similarity. Contrary to the traditional zero-shot learning approaches that are built upon attribute presence, our approach bypasses the laborious attribute-class relation annotations for unseen classes. In addition, our proposed approach renders text-only training possible, hence, the training can be augmented without the need to collect additional image data. The experimental results show that our method yields state-of-the-art results for unsupervised ZSL in three benchmark datasets.

****

Areas of Attention for Image Captioning

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, Jakob Verbeek; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1242-1250

We propose "Areas of Attention", a novel attention-based model for automatic image captioning. Our approach models the dependencies between image regions, caption words, and the state of an RNN language model, using three pairwise interactions. In contrast to previous attention-based approaches that associate image regions to the RNN state, our method allows a direct association between caption words and image regions. During training these associations are inferred from image-level captions, akin to weakly-supervised object detector training. These associations help to improve captioning by localizing the corresponding regions during testing. We also propose and compare different ways of generating attention areas: CNN activation grids, object proposals, and spatial transformers nets applied in a convolutional fashion. Spatial transformers give the best results, since they allow for image specific attention areas, and can be trained jointly with the rest of the network. Our attention mechanism and spatial transformer attention areas together yield state-of-the-art results on the MSCOCO dataset.

****

Generative Modeling of Audible Shapes for Object Perception

Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, William T. Freeman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1251-1260

Humans infer rich knowledge of objects from both auditory and visual cues. Building a machine of such competency, however, is very challenging, due to the great difficulty in capturing large-scale, clean data of objects with both their appearance and the sound they make. In this paper, we present a novel, open-source pipeline that generates audio-visual data, purely from 3D object shapes and their physical properties. Through comparison with audio recordings and human behavioral studies, we validate the accuracy of the sounds it generates. Using this generative model, we are able to construct a synthetic audio-visual dataset, namely Sound-20K, for object perception tasks. We demonstrate that auditory and visual information play complementary roles in object perception, and further, that the representation learned on synthetic audio-visual data can transfer to real-world scenarios.

****

Scene Graph Generation From Objects, Phrases and Region Captions

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1261-1270

Object detection, scene graph generation and region captioning, which are three

scene understanding tasks at different semantic levels, are tied together: scene graphs are generated on top of objects detected in an image with their pairwise relationship predicted, while region captioning gives a language description of the objects, their attributes, relations and other context information. In this work, to leverage the mutual connections across semantic levels, we propose a novel neural network model, termed as Multi-level Scene Description Network (denoted as MSDN), to solve the three vision tasks jointly in an end-to-end manner. Object, phrase, and caption regions are first aligned with a dynamic graph based on their spatial and semantic connections. Then a feature refining structure is used to pass messages across the three levels of semantic tasks through the graph. We benchmark the learned model on three tasks, and show the joint learning across three tasks with our proposed method can bring mutual improvements over previous models. Particularly, on the scene graph generation task, our proposed method outperforms the state-of-art method with more than 3% margin.
****
Recurrent Multimodal Interaction for Referring Image Segmentation
Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1271-1280
In this paper we are interested in the problem of image segmentation given natural language descriptions, i.e. referring expressions. Existing works tackle this problem by first modeling images and sentences independently and then segment images by combining these two types of representations. We argue that learning word-to-image interaction is more native in the sense of jointly modeling two modalities for the image segmentation task, and we propose convolutional multimodal LSTM to encode the sequential interactions between individual words, visual information, and spatial information. We show that our proposed model outperforms the baseline model on benchmark datasets. In addition, we analyze the intermediate output of the proposed multimodal LSTM approach and empirically explain how this approach enforces a more effective word-to-image interaction.
****
Learning Feature Pyramids for Human Pose Estimation
Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1281-1290
Articulated human pose estimation is a fundamental yet challenging task in computer vision. The difficulty is particularly pronounced in scale variations of human body parts when camera view changes or severe foreshortening happens. Although pyramid methods are widely used to handle scale changes at inference time, learning feature pyramids in deep convolutional neural networks (DCNNs) is still not well explored. In this work, we design a Pyramid Residual Module (PRMs) to enhance the invariance in scales of DCNNs. Given input features, the PRMs learn convolutional filters on various scales of input features, which are obtained with different subsampling ratios in a multi-branch network. Moreover, we observe that it is inappropriate to adopt existing methods to initialize the weights of multi-branch networks, which achieve superior performance than plain networks in many tasks recently. Therefore, we provide theoretic derivation to extend the current weight initialization scheme to multi-branch network structures. We investigate our method on two standard benchmarks for human pose estimation. Our approach obtains state-of-the-art results on both benchmarks. Code is available at https://github.com/bearpaw/PyraNet.
****
Structured Attentions for Visual Question Answering
Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, Yi Ma; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1291-1300
Visual attention, which assigns weights to image regions according to their relevance to a question, is considered as an indispensable part by most Visual Question Answering models. Although the questions may involve complex relations among multiple regions, few attention models can effectively encode such cross-region relations. In this paper,we emonstrate the importance of encoding such relations by showing the limited effective receptive field of ResNet on two datasets, an

d propose to model the visual attention as a multivariate distribution over a grid-structured Conditional Random Field on image regions. We demonstrate how to convert the iterative inference algorithms, Mean Field and Loopy Belief Propagation, as recurrent layers of an end-to-end neural network. We empirically evaluated our model on 3 datasets, in which it surpasses the best baseline model of the newly released CLEVR dataset by 9.5%, and the best published model on the VQA dataset by 1.25%. Source code is available at https://github.com/zhuchen03/vqa-sva.

****

Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection

Debidatta Dwibedi, Ishan Misra, Martial Hebert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1301-1310

A major impediment in rapidly deploying object detection models for instance detection is the lack of large annotated datasets. For example, finding a large labeled dataset containing instances in a particular kitchen is unlikely. Each new environment with new instances requires expensive data collection and annotation. In this paper, we propose a simple approach to generate large annotated instance datasets with minimal effort. Our key insight is that ensuring only patch-level realism provides enough training signal for current object detector models. We automatically `cut' object instances and `paste' them on random backgrounds. A naive way to do this results in pixel artifacts which result in poor performance for trained models. We show how to make detectors ignore these artifacts during training and generate data that gives competitive performance on real data. Our method outperforms existing synthesis approaches and when combined with real images improves relative performance by more than 21% on benchmark datasets. In a cross-domain setting, our synthetic data combined with just 10% real data outperforms models trained on all real data.

****

Cascaded Feature Network for Semantic Segmentation of RGB-D Images

Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, Hui Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1311-1319

Fully convolutional network (FCN) has been successfully applied in semantic segmentation of scenes represented with RGB images. Images augmented with depth channel provide more understanding of the geometric information of the scene in the image. The question is how to best exploit this additional information to improve the segmentation performance. In this paper, we present a neural network with multiple branches for segmenting RGB-D images. Our approach is to use the available depth to split the image into layers with common visual characteristic of objects/scenes, or common "scene-resolution". We introduce context-aware receptive field (CaRF) which provides a better control on the relevant contextual information of the learned features. Equipped with CaRF, each branch of the network semantically segments relevant similar scene-resolution, leading to a more focused domain which is easier to learn. Furthermore, our network is cascaded with features from one branch augmenting the features of adjacent branch. We show that such cascading of features enriches the contextual information of each branch and enhances the overall performance. The accuracy that our network achieves outperforms the state-of-the-art methods on two public datasets.

****

Encoder Based Lifelong Learning

Amal Rannen, Rahaf Aljundi, Matthew B. Blaschko, Tinne Tuytelaars; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1320-1328

This paper introduces a new lifelong learning solution where a single model is trained for a sequence of tasks. The main challenge that vision systems face in this context is catastrophic forgetting: as they tend to adapt to the most recently seen task, they lose performance on the tasks that were learned previously. Our method aims at preserving the knowledge of the previous tasks while learning a new one by using autoencoders. For each task, an under-complete autoencoder is learned, capturing the features that are crucial for its achievement. When a ne w task is presented to the system, we prevent the reconstructions of the features with these autoencoders from changing, which has the effect of preserving the information on which the previous tasks are mainly relying. At the same time, the features are given space to adjust to the most recent environment as only their projection into a low dimension submanifold is controlled. The proposed system is evaluated on image classification tasks and shows a reduction of forgetting over the state-of-the-art.

****

Transitive Invariance for Self-Supervised Visual Representation Learning
Xiaolong Wang, Kaiming He, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1329-1338

Learning visual representations with self-supervised learning has become popular in computer vision. The idea is to design auxiliary tasks where labels are free to obtain. Most of these tasks end up providing data to learn specific kinds of invariance useful for recognition. In this paper, we propose to exploit different self-supervised approaches to learn representations invariant to (i) inter-instance variations (two objects in the same class should have similar features) and (ii) intra-instance variations (viewpoint, pose, deformations, illumination, etc). Instead of combining two approaches with multi-task learning, we argue to organize and reason the data with multiple variations. Specifically, we propose to generate a graph with millions of objects mined from hundreds of thousands of videos. The objects are connected by two types of edges which correspond to two types of invariance: "different instances but a similar viewpoint and category" and "different viewpoints of the same instance". By applying simple transitivity on the graph with these edges, we can obtain pairs of images exhibiting richer visual invariance. We use this data to train a Triplet-Siamese network with VGG16 as the base architecture and apply the learned representations to different recognition tasks. For object detection, we achieve 63.2% mAP on PASCAL VOC 2007 using Fast R-CNN (compare to 67.3% with ImageNet pre-training). For the challenging COCO dataset, our method is surprisingly close (23.5%) to the ImageNet-supervised counterpart (24.4%) using the Faster R-CNN framework. We also show that our network can perform significantly better than the ImageNet network in the surface normal estimation task.

****

Weakly Supervised Learning of Deep Metrics for Stereo Reconstruction
Stepan Tulyakov, Anton Ivanov, Francois Fleuret; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1339-1348

Deep-learning metrics have recently demonstrated extremely good performance to match image patches for stereo reconstruction. However, training such metrics requires large amount of labeled stereo images, which can be difficult or costly to collect for certain applications (consider for example satellite stereo imaging). Moreover, labels from the depth sensors are often noisy. The main contribution of our work is a new weakly-supervised method for learning deep metrics from unlabeled stereo images, given coarse information about the scenes and the optical system. Our method alternatively optimizes the metric with a standard stochastic gradient descent, and applies stereo constraints to regularize its prediction. Experiments on reference data-sets show that, for a given network architecture, training with this new method without ground-truth produces a metric with performance as good as state-of-the-art baselines trained with the said ground-truth. This work has three practical implications. Firstly, it helps to overcome limitations of training sets, in particular noisy ground truth. Secondly it allows to use much more training data during learning. Thirdly, it allows to tune deep metric for a particular stereo system, even if ground truth is not available.

****

Fine-Grained Recognition in the Wild: A Multi-Task Domain Adaptation Approach
Timnit Gebru, Judy Hoffman, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1349-1358

While fine-grained object recognition is an important problem in computer vision, current models are unlikely to accurately classify objects in the wild. These fully supervised models need additional annotated images to classify objects in every new scenario, a task that is infeasible. However, sources such as e-commerce websites and field guides provide annotated images for many classes. In this work, we study fine-grained domain adaptation as a step towards overcoming the dataset shift between easily acquired annotated images and the real world. Adaptation has not been studied in the fine-grained setting where annotations such as attributes could be used to increase performance. Our work uses an attribute based multitask adaptaion loss to increase accuracy from a baseline of 3.4% to 19% in the semi-supervised adaptation case. Prior domain adaptation works have been benchmarked on small datasets such as [45] with a total of 795 images for some domains, or simplistic datasets such as [40] consisting of digits. We perform experiments on a new challenging fine-grained dataset of cars consisting of 1, 095, 021 images of 2, 657 categories of cars drawn from e-commerce websites and Google Street View.

****

SORT: Second-Order Response Transform for Visual Recognition
Yan Wang, Lingxi Xie, Chenxi Liu, Siyuan Qiao, Ya Zhang, Wenjun Zhang, Qi Tian, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1359-1368

In this paper, we reveal the importance and benefits of introducing second-order operations into deep neural networks. We propose a novel approach named Second-Order Response Transform (SORT), which appends element-wise product transform to the linear sum of a two-branch network module. A direct advantage of SORT is to facilitate cross-branch response propagation, so that each branch can update its weights based on the current status of the other branch. Moreover, SORT augments the family of transform operations and increases the nonlinearity of the network, making it possible to learn flexible functions to fit the complicated distribution of feature space. SORT can be applied to a wide range of network architectures, including a branched variant of a chain-styled network and a residual network, with very light-weighted modifications. We observe consistent accuracy gain on both small (CIFAR10, CIFAR100 and SVHN) and big (ILSVRC2012) datasets. In addition, SORT is very efficient, as the extra computation overhead is less than 5%.

****

Adversarial Examples for Semantic Segmentation and Object Detection
Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1369-1378

It has been well demonstrated that adversarial examples, i.e., natural images with visually imperceptible perturbations added, cause deep networks to fail on image classification. In this paper, we extend adversarial examples to semantic segmentation and object detection which are much more difficult. Our observation is that both segmentation and detection are based on classifying multiple targets on an image (e.g., the target is a pixel or a receptive field in segmentation, and an object proposal in detection). This inspires us to optimize a loss function over a set of targets for generating adversarial perturbations. Based on this, we propose a novel algorithm named Dense Adversary Generation (DAG), which applies to the state-of-the-art networks for segmentation and detection. We find that the adversarial perturbations can be transferred across networks with different training data, based on different architectures, and even for different recognition tasks. In particular, the transfer ability across networks with the same architecture is more significant than in other cases. Besides, we show that summing up heterogeneous perturbations often leads to better transfer performance, which provides an effective method of black-box adversarial attack.

****

Genetic CNN
Lingxi Xie, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1379-1388

The deep convolutional neural network (CNN) is the state-of-the-art solution for large-scale visual recognition. Following some basic principles such as increasing network depth and constructing highway connections, researchers have manuall y designed a lot of fixed network architectures and verified their effectiveness. In this paper, we discuss the possibility of learning deep network structures automatically. Note that the number of possible network structures increases exponentially with the number of layers in the network, which motivates us to adopt the genetic algorithm to efficiently explore this large search space. The core idea is to propose an encoding method to represent each network structure in a fixed-length binary string. The genetic algorithm is initialized by generating a set of randomized individuals. In each generation, we define standard genetic operations, e.g., selection, mutation and crossover, to generate competitive individuals and eliminate weak ones. The competitiveness of each individual is defined as its recognition accuracy, which is obtained via a standalone training process on a reference dataset. We run the genetic process on CIFAR10, a small-scale dataset, demonstrating its ability to find high-quality structures which are little studied before. The learned powerful structures are also transferrable to the ILSVRC2012 dataset for large-scale visual recognition.

****

Channel Pruning for Accelerating Very Deep Neural Networks

Yihui He, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1389-1397

In this paper, we introduce a new channel pruning method to accelerate very deep convolutional neural networks.Given a trained CNN model, we propose an iterative two-step algorithm to effectively prune each layer, by a LASSO regression based channel selection and least square reconstruction. We further generalize this algorithm to multi-layer and multi-branch cases. Our method reduces the accumulated error and enhance the compatibility with various architectures. Our pruned VGG-16 achieves the state-of-the-art results by 5x speed-up along with only 0.3% increase of error. More importantly, our method is able to accelerate modern networks like ResNet, Xception and suffers only 1.4%, 1.0% accuracy loss under 2x speed-up respectively, which is significant.

****

Infinite Latent Feature Selection: A Probabilistic Latent Graph-Based Ranking Approach

Giorgio Roffo, Simone Melzi, Umberto Castellani, Alessandro Vinciarelli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1398-1406

Feature selection is playing an increasingly significant role with respect to many computer vision applications spanning from object recognition to visual object tracking. However, most of the recent solutions in feature selection are not robust across different and heterogeneous set of data. In this paper, we address this issue proposing a robust probabilistic latent graph-based feature selection algorithm that performs the ranking step while considering all the possible subsets of features, as paths on a graph, bypassing the combinatorial problem analytically. An appealing characteristic of the approach is that it aims to discover an abstraction behind low-level sensory data, that is, relevancy. Relevancy is modelled as a latent variable in a PLSA-inspired generative process that allows the investigation of the importance of a feature when injected into an arbitrary set of cues. The proposed method has been tested on ten diverse benchmarks, and compared against eleven state of the art feature selection methods. Results show that the proposed approach attains the highest performance levels across many different scenarios and difficulties, thereby confirming its strong robustness while setting a new state of the art in feature selection domain.

****

Video Fill in the Blank Using LR/RL LSTMs With Spatial-Temporal Attentions

Amir Mazaheri, Dong Zhang, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1407-1416

Given a video and a description sentence with one missing word, "source sentence", Video-Fill-In-the-Blank (VFIB) problem is to find the missing word automatically. The contextual information of the sentence, as well as visual cues from the video, are important to infer the missing word accurately. Since the source sentence is broken into two fragments: the sentence's left fragment (before the bla nk) and the sentence's right fragment (after the blank), traditional Recurrent Neural Networks cannot encode this structure accurately because of many possible variations of the missing word in terms of the location and type of the word in the source sentence. For example, a missing word can be the first word or be in the middle of the sentence and it can be a verb or an adjective. In this paper, we propose a framework to tackle the textual encoding: Two separate LSTMs (the LR and RL LSTMs) are employed to encode the left and right sentence fragments and a novel structure is introduced to combine each fragment with an "external memory" corresponding to the opposite fragments. For the visual encoding, end-to-end spatial and temporal attention models are employed to select discriminative visual representations to find the missing word. In the experiments, we demonstrate the superior performance of the proposed method on challenging VFIB problem. Furthermore, we introduce an extended and more generalized version of VFIB, which is not limited to a single blank. Our experiments indicate the generalization capability of our method in dealing with such more realistic scenarios.

****

Primary Video Object Segmentation via Complementary CNNs and Neighborhood Reversible Flow

Jia Li, Anlin Zheng, Xiaowu Chen, Bin Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1417-1425

This paper proposes a novel approach for segmenting primary video objects by using Complementary Convolutional Neural Networks (CCNN) and neighborhood reversible flow. The proposed approach first pre-trains CCNN on massive images with manually annotated salient objects in an end-to-end manner, and the trained CCNN has two separate branches that simultaneously handle two complementary tasks, i.e., foregroundness and backgroundness estimation. By applying CCNN on each video frame, the spatial foregroundness and backgroundness maps can be initialized, which are then propagated between various frames so as to segment primary video objects and suppress distractors. To enforce efficient temporal propagation, we divide each frame into superpixels and construct neighborhood reversible flow that reflects the most reliable temporal correspondences between superpixels in far-away frames. Within such flow, the initialized foregroundness and backgroundness can be efficiently and accurately propagated along the temporal axis so that primary video objects gradually pop-out and distractors are well suppressed. Extensive experimental results on three video datasets show that the proposed approach achieves impressive performance in comparisons with 18 state-of-the-art models.

****

Attentive Semantic Video Generation Using Captions

Tanya Marwah, Gaurav Mittal, Vineeth N. Balasubramanian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1426-1434

This paper proposes a network architecture to perform variable length semantic video generation using captions. We adopt a new perspective towards video generation where we allow the captions to be combined with the long-term and short-term dependencies between video frames and thus generate a video in an incremental manner. Our experiments demonstrate our network architecture's ability to distinguish between objects, actions and interactions in a video and combine them to generate videos for unseen captions. The network also exhibits the capability to perform spatio-temporal style transfer when asked to generate videos for a sequence of captions. We also show that the network's ability to learn a latent representation allows it generate videos in an unsupervised manner and perform other tasks such as action recognition.

****

Following Gaze in Video

Adria Recasens, Carl Vondrick, Aditya Khosla, Antonio Torralba; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1435-1443

Following the gaze of people inside videos is an important signal for understanding people and their actions. In this paper, we present an approach for following gaze in video by predicting where a person (in the video) is looking even when the object is in a different frame. We collect VideoGaze, a new dataset which we use as a benchmark to both train and evaluate models. Given one frame with a p

erson in it, our model estimates a density for gaze location in every frame and the probability that the person is looking in that particular frame. A key aspect of our approach is an end-to-end model that jointly estimates: saliency, gaze pose, and geometric relationships between views while only using gaze as supervision. Visualizations suggest that the model learns to internally solve these intermediate tasks automatically without additional supervision. Experiments show that our approach follows gaze in video better than existing approaches, enabling a richer understanding of human activities in video.
****

Adaptive RNN Tree for Large-Scale Human Action Recognition
Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, Siwei Lyu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1444-1452
In this work, we present the RNN Tree (RNN-T), an adaptive learning framework for skeleton based human action recognition. Our method categorizes action classes and uses multiple Recurrent Neural Networks (RNNs) in a tree-like hierarchy. The RNNs in RNN-T are co-trained with the action category hierarchy, which determines the structure of RNN-T. Actions in skeletal representations are recognized via a hierarchical inference process, during which individual RNNs differentiate finer-grained action classes with increasing confidence. Inference in RNN-T ends when any RNN in the tree recognizes the action with high confidence, or a leaf node is reached. RNN-T effectively addresses two main challenges of large-scale action recognition: (i) able to distinguish fine-grained action classes that are intractable using a single network, and (ii) adaptive to new action classes by augmenting an existing model. We demonstrate the effectiveness of RNN-T/ACH method and compare it with the state-of-the-art methods on a large-scale dataset and several existing benchmarks.
****

Spatio-Temporal Person Retrieval via Natural Language Queries
Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1453-1462
In this paper, we address the problem of spatio-temporal person retrieval from videos using a natural language query, in which we output a tube (i.e., a sequence of bounding boxes) which encloses the person described by the query. For this problem, we introduce a novel dataset consisting of videos containing people annotated with bounding boxes for each second and with five natural language descriptions. To retrieve the tube of the person described by a given natural language query, we design a model that combines methods for spatio-temporal human detection and multimodal retrieval. We conduct comprehensive experiments to compare a variety of tube and text representations and multimodal retrieval methods, and present a strong baseline in this task as well as demonstrate the efficacy of our tube representation and multimodal feature embedding technique. Finally, we demonstrate the versatility of our model by applying it to two other important tasks.
****

Automatic Spatially-Aware Fashion Concept Discovery
Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1463-1471
This paper proposes an automatic spatially-aware concept discovery approach using weakly labeled image-text data from shopping websites. We first fine-tune GoogleNet by jointly modeling clothing images and their corresponding descriptions in a visual-semantic embedding space. Then, for each attribute (word), we generate its spatially-aware representation by combining its semantic word vector representation with its spatial representation derived from the convolutional maps of the fine-tuned network. The resulting spatially-aware representations are further used to cluster attributes into multiple groups to form spatially-aware concepts (e.g., the neckline concept might consist of attributes like v-neck, round-neck, etc). Finally, we decompose the visual-semantic embedding space into multiple concept-specific subspaces, which facilitates structured browsing and attribu

te-feedback product retrieval by exploiting multimodal linguistic regularities. We conducted extensive experiments on our newly collected Fashion200K dataset, and results on clustering quality evaluation and attribute-feedback product retrieval task demonstrate the effectiveness of our automatically discovered spatially-aware concepts.

****

ChromaTag: A Colored Marker and Fast Detection Algorithm

Joseph DeGol, Timothy Bretl, Derek Hoiem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1472-1481

Current fiducial marker detection algorithms rely on marker IDs for false positive rejection. Time is wasted on potential detections that will eventually be rejected as false positives. We introduce ChromaTag, a fiducial marker and detection algorithm designed to use opponent colors to limit and quickly reject initial false detections and grayscale for precise localization. Through experiments, we show that ChromaTag is significantly faster than current fiducial markers while achieving similar or better detection accuracy. We also show how tag size and viewing direction effect detection accuracy. Our contribution is significant because fiducial markers are often used in real-time applications (e.g. marker assisted robot navigation) where heavy computation is required by other parts of the system.

****

Adversarial Image Perturbation for Privacy Protection -- A Game Theory Perspective

Seong Joon Oh, Mario Fritz, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1482-1491

Users like sharing personal photos with others through social media. At the same time, they might want to make automatic identification in such photos difficult or even impossible. Classic obfuscation methods such as blurring are not only unpleasant but also not as effective as one would expect. Recent studies on adversarial image perturbations (AIP) suggest that it is possible to confuse recognition systems effectively without unpleasant artifacts. However, in the presence of counter measures against AIPs, it is unclear how effective AIP would be in particular when the choice of counter measure is unknown. Game theory provides tools for studying the interaction between agents with uncertainties in the strategies. We introduce a general game theoretical framework for the user-recogniser dynamics, and present a case study that involves current state of the art AIP and person recognition techniques. We derive the optimal strategy for the user that assures an upper bound on the recognition rate independent of the recogniser's counter measure. Code is available at https://goo.gl/hgvbNK.

****

WeText: Scene Text Detection Under Weak Supervision

Shangxuan Tian, Shijian Lu, Chongshou Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1492-1500

The requiring of large amounts of annotated training data has become a common constraint on various deep learning systems. In this paper, we propose a weakly supervised scene text detection method (WeText) that trains robust and accurate scene text detection models by learning from unannotated or weakly annotated data. With a "light" supervised model trained on a small fully annotated dataset, we explore semi-supervised and weakly supervised learning on a large unannotated dataset and a large weakly annotated dataset, respectively. For the unsupervised learning, the light supervised model is applied to the unannotated dataset to search for more character training samples, which are further combined with the small annotated dataset to retrain a superior character detection model. For the weakly supervised learning, the character searching is guided by high-level annotations of words/text lines that are widely available and also much easier to prepare. In addition, we design an unified scene character detector by adapting regression based deep networks, which greatly relieves the error accumulation issue that widely exists in most traditional approaches. Extensive experiments across different unannotated and weakly annotated datasets show that the scene text detection performance can be clearly boosted under both scenarios, where the weakly supervised learning can achieve the state-of-the-art performance by using only 229 fully annotated scene text images.

****

Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization

Xun Huang, Serge Belongie; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1501-1510

Gatys et al. recently introduced a neural algorithm that renders a content image in the style of another image, achieving so-called style transfer. However, their framework requires a slow iterative optimization process, which limits its practical application. Fast approximations with feed-forward neural networks have been proposed to speed up neural style transfer. Unfortunately, the speed improvement comes at a cost: the network is usually tied to a fixed set of styles and cannot adapt to arbitrary new styles. In this paper, we present a simple yet effective approach that for the first time enables arbitrary style transfer in real-time. At the heart of our method is a novel adaptive instance normalization (AdaIN) layer that aligns the mean and variance of the content features with those of the style features. Our method achieves speed comparable to the fastest existing approach, without the restriction to a pre-defined set of styles. In addition, our approach allows flexible user controls such as content-style trade-off, style interpolation, color & spatial controls, all using a single feed-forward neural network.

****

Photographic Image Synthesis With Cascaded Refinement Networks

Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1511-1520

We present an approach to synthesizing photographic images conditioned on semantic layouts. Given a semantic label map, our approach produces an image with photographic appearance that conforms to the input layout. The approach thus functions as a rendering engine that takes a two-dimensional semantic specification of the scene and produces a corresponding photographic image. Unlike recent and contemporaneous work, our approach does not rely on adversarial training. We show that photographic images can be synthesized from semantic layouts by a single feedforward network with appropriate structure, trained end-to-end with a direct regression objective. The presented approach scales seamlessly to high resolutions; we demonstrate this by synthesizing photographic images at 2-megapixel resolution, the full resolution of our training data. Extensive perceptual experiments on datasets of outdoor and indoor scenes demonstrate that images synthesized by the presented approach are considerably more realistic than alternative approaches.

****

SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again

Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, Nassir Navab; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1521-1529

We present a novel method for detecting 3D model instances and estimating their 6D poses from RGB data in a single shot. To this end, we extend the popular SSD paradigm to cover the full 6D pose space and train on synthetic model data only. Our approach competes or surpasses current state-of-the-art methods that leverage RGB-D data on multiple challenging datasets. Furthermore, our method produces these results at around 10Hz, which is many times faster than the related methods. For the sake of reproducibility, we make our trained networks and detection code publicly available.

****

Unsupervised Creation of Parameterized Avatars

Lior Wolf, Yaniv Taigman, Adam Polyak; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1530-1538

We study the problem of mapping an input image to a tied pair consisting of a vector of parameters and an image that is created using a graphical engine from the vector of parameters. The mapping's objective is to have the output image as similar as possible to the input image. During training, no supervision is given in the form of matching inputs and outputs. This learning problem extends two literature problems: unsupervised domain adaptation and cross domain transfer. We define a generalization bound that is based on discrepancy, and employ a GAN to implement a network solution that corresponds to this bound. Experimentally, our method is shown to solve the problem of automatically creating avatars.

****

Learning for Active 3D Mapping

Karel Zimmermann, Tomas Petricek, Vojtech Salansky, Tomas Svoboda; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1539-1547

We propose an active 3D mapping method for depth sensors, which allow individual control of depth-measuring rays, such as the newly emerging Solid State Lidars. The method simultaneously (i) learns to reconstruct a dense 3D voxel-map from sparse depth measurements, and (ii) optimizes the reactive control of depth-measuring rays. To make the first step towards the online control optimization, we propose a fast greedy algorithm, which needs to update its cost function in only a small fraction of possible rays. The approximation ratio of the greedy algorithm is derived. Experimental evaluation on the subset of the Kitti dataset demonstrates significant improvement in the 3D map accuracy when learning-to-reconstruct from sparse measurements is coupled with the optimization where-to-measure.

****

Toward Perceptually-Consistent Stereo: A Scanline Study

Jialiang Wang, Daniel Glasner, Todd Zickler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1548-1556

Two types of information exist in a stereo pair: correlation (matching) and decorrelation (half-occlusion). Vision science has shown that both types of information are used in the visual cortex, and that people can perceive depth even when correlation cues are absent or very weak, a capability that remains absent from most computational stereo systems. As a step toward stereo algorithms that are more consistent with these perceptual phenomena, we re-examine the topic of scanline stereo as energy minimization. We represent a disparity profile as a piecewise smooth function with explicit breakpoints between its smooth pieces, and we show this allows correlation and decorrelation to be integrated into an objective that requires only two types of local information: the correlation and its spatial gradient. Experimentally, we show the global optimum of this objective matches human perception on a broad collection of wellknown perceptual stimuli, and that it also provides reasonable piecewise-smooth interpretations of depth in natural images, even without exploiting monocular boundary cues.

****

Surface Normals in the Wild

Weifeng Chen, Donglai Xiang, Jia Deng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1557-1566

We study the problem of single-image depth estimation for images in the wild. We collect human annotated surface normals and use them to help train a neural network that directly predicts pixel-wise depth. We propose two novel loss functions for training with surface normal annotations. Experiments on NYU Depth, KITTI, and our own dataset demonstrate that our approach can significantly improve the quality of depth estimation in the wild.

****

Unsupervised Learning of Stereo Matching

Chao Zhou, Hong Zhang, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1567-1575

In recent years, convolutional neural networks have shown its strong power for stereo matching cost learning. Current approaches learn the parameters of their models from public datasets with ground truth disparity. However, due to the limitations of these datasets and the difficulty of collecting new stereo data, current methods fail in real-life cases. In this work, we present a framework for learning stereo matching cost without human supervision. Our method updates the network parameter in a iterative manner. It starts with randomly initialized network. Correct matchings are carefully picked and used as training data in each rou nd. In the end, the networks converges to a stable state, which performs comparably with supervised methods on various benchmarks.

****

Unrestricted Facial Geometry Reconstruction Using Image-To-Image Translation

Matan Sela, Elad Richardson, Ron Kimmel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1576-1585

It has been recently shown that neural networks can recover the geometric structure of a face from a single given image. A common denominator of most existing face geometry reconstruction methods is the restriction of the solution space to some low-dimensional subspace. While such a model significantly simplifies the reconstruction problem, it is inherently limited in its expressiveness. As an alternative, we propose an Image-to-Image translation network that jointly maps the input image to a depth image and a facial correspondence map. This explicit pixel-based mapping can then be utilized to provide high quality reconstructions of diverse faces under extreme expressions, using a purely geometric refinement process. In the spirit of recent approaches, the network is trained only with synthetic data, and is then evaluated on in-the-wild facial images. Both qualitative and quantitative analyses demonstrate the accuracy and the robustness of our approach.

****

Learned Multi-Patch Similarity

Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, Konrad Schindler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1586-1594

Estimating a depth map from multiple views of a scene is a fundamental task in computer vision. As soon as more than two viewpoints are available, one faces the very basic question how to measure similarity across >2 image patches. Surprisingly, no direct solution exists, instead it is common to fall back to more or less robust averaging of two-view similarities. Encouraged by the success of machine learning, and in particular convolutional neural networks, we propose to learn a matching function which directly maps multiple image patches to a scalar similarity score. Experiments on several multi-view datasets demonstrate that this approach has advantages over methods based on pairwise patch similarity.

****

Click Here: Human-Localized Keypoints as Guidance for Viewpoint Estimation

Ryan Szeto, Jason J. Corso; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1595-1604

We motivate and address a human-in-the-loop variant of the monocular viewpoint estimation task in which the location and class of one semantic object keypoint is available at test time. In order to leverage the keypoint information, we devise a Convolutional Neural Network called Click-Here CNN (CH-CNN) that integrates the keypoint information with activations from the layers that process the image. It transforms the keypoint information into a 2D map that can be used to weigh features from certain parts of the image more heavily. The weighted sum of these spatial features is combined with global image features to provide relevant information to the prediction layers. To train our network, we collect a novel dataset of 3D keypoint annotations on thousands of CAD models, and synthetically render millions of images with 2D keypoint information. On test instances from PASCAL 3D+, our model achieves a mean class accuracy of 90.7%, whereas the state-of-the-art baseline only obtains 85.7% mean class accuracy, justifying our argument for human-in-the-loop inference.

****

Unsupervised Adaptation for Deep Stereo

Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, Luigi Di Stefano; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1605-1613

Recent ground-breaking works have shown that deep neural networks can be trained end-to-end to regress dense disparity maps directly from image pairs. Computer generated imagery is deployed to gather the large data corpus required to train such networks, an additional fine-tuning allowing to adapt the model to work wel l also on real and possibly diverse environments. Yet, besides a few public datasets such as Kitti, the ground-truth needed to adapt the network to a new scenario is hardly available in practice. In this paper we propose a novel unsupervised adaptation approach that enables to fine-tune a deep learning stereo model without any ground-truth information. We rely on off-the-shelf stereo algorithms together with state-of-the-art confidence measures, the latter able to ascertain upon correctness of the measurements yielded by former. Thus, we train the network based on a novel loss-function that penalizes predictions disagreeing with the highly confident disparities provided by the algorithm and enforces a smoothness constraint. Experiments on popular datasets (KITTI 2012, KITTI 2015 and Middlebury 2014) and other challenging test images demonstrate the effectiveness of our proposal.

****

Composite Focus Measure for High Quality Depth Maps

Parikshit Sakurikar, P. J. Narayanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1614-1622

Depth from focus is a highly accessible method to estimate the 3D structure of everyday scenes. Today's DSLR and mobile cameras facilitate the easy capture of multiple focused images of a scene. Focus measures (FMs) that estimate the amount of focus at each pixel form the basis of depth-from-focus methods. Several FMs have been proposed in the past and new ones will emerge in the future, each with their own strengths. We estimate a weighted combination of standard FMs that outperforms others on a wide range of scene types. The resulting composite focus measure consists of FMs that are in consensus with one another but not in chorus. Our two-stage pipeline first estimates fine depth at each pixel using the composite focus measure. A cost-volume propagation step then assigns depths from confident pixels to others. We can generate high quality depth maps using just the top five FMs from our composite focus measure. This is a positive step towards depth estimation of everyday scenes with no special equipment.

****

Reconstruction-Based Disentanglement for Pose-Invariant Face Recognition

Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N. Metaxas, Manmohan Chandraker; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1623-1632

Deep neural networks (DNNs) trained on large-scale datasets have recently achieved impressive improvements in face recognition. But a persistent challenge remains to develop methods capable of handling large pose variations that are relatively under-represented in training data. This paper presents a method for learning a feature representation that is invariant to pose, without requiring extensive pose coverage in training data. We first propose to generate non-frontal views from a single frontal face, in order to increase the diversity of training data while preserving accurate facial details that are critical for identity discrimination. Our next contribution is to seek a rich embedding that encodes identity features, as well as non-identity ones such as pose and landmark locations. Finally, we propose a new feature reconstruction metric learning to explicitly disentangle identity and pose, by demanding alignment between the feature reconstructions through various combinations of identity and pose features, which is obtained from two images of the same subject. Experiments on both controlled and in-the-wild face datasets, such as MultiPIE, 300WLP and the profile view database CFP, show that our method consistently outperforms the state-of-the-art, especially on images with large head pose variations.

****

Recurrent 3D-2D Dual Learning for Large-Pose Facial Landmark Detection

Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, Ashraf Kassim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1633-1642

Despite remarkable progress of face analysis techniques, detecting landmarks on large-pose faces is still difficult due to self-occlusion, subtle landmark difference and incomplete information. To address these challenging issues, we introduce a novel recurrent 3D-2D dual learning model that alternatively performs 2D-b ased 3D face model refinement and 3D-to-2D projection based 2D landmark refinement to reliably reason about self-occluded landmarks, precisely capture the subtle landmark displacement and accurately detect landmarks even in presence of extremely large poses. The proposed model presents the first loop-closed learning framework that effectively exploits the informative feedback from the 3D-2D learning and its dual 2D-3D refinement tasks in a recurrent manner. Benefiting from these two mutual-boosting steps, our proposed model demonstrates appealing robustness to large poses (up to profile pose) and outstanding ability to capture fine-scale landmark displacement compared with existing 3D models. It achieves new state-of-the-art on the challenging AFLW benchmark. Moreover, our proposed model introduces a new architectural design that economically utilizes intermediate features and achieves 4x faster speed than its deep learning based counterparts.

****

Anchored Regression Networks Applied to Age Estimation and Super Resolution
Eirikur Agustsson, Radu Timofte, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1643-1652

We propose the Anchored Regression Network (ARN), a nonlinear regression network which can be seamlessly integrated into various networks or can be used stand-alone when the features have already been fixed. Our ARN is a smoothed relaxation of a piecewise linear regressor through the combination of multiple linear regressors over soft assignments to anchor points. When the anchor points are fixed the optimal ARN regressors can be obtained with a closed form global solution, otherwise ARN admits end-to-end learning with standard gradient based methods. We demonstrate the power of the ARN by applying it to two very diverse and challenging tasks: age prediction from face images and image super-resolution. In both cases, ARNs yield strong results.

****

Infant Footprint Recognition
Eryun Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1653-1660

Infant recognition has received increasing attention in recent years in many applications, such as tracking child vaccination and identifying missing children. Due to the lack of efficient identification methods for infants and newborns, the current methods of infant recognition rely on identification of parents or certificates of identity. While biometric recognition technologies (e.g., face and fingerprint recognition) have been widely deployed in many applications for recognizing adults and teenagers, no such recognition systems yet exist for infants or newborns. One of the major problems is that the biometric traits of infants and newborns are either not permanent (e.g., face) or difficult to capture (e.g., fingerprint) due to lack of appropriate sensors. In this paper, we investigate the feasibility of infant recognition by their footprint using a 500 ppi commodity friction ridge sensor. We collected an infant footprint dataset in three sessions, consisting of 60 subjects, with age range from 1 to 9 months. We proposed a new minutia descriptor based on deep convolutional neural network for measuring minutiae similarity. The descriptor is compact and highly discriminative. We conducted verification experiments for both single enrolled template and fusion of multiple enrolled templates, and show the impact of age and time gap on matching performance. Comparison experiments with state of the art algorithm show the advantage of the proposed minutia descriptor.

****

Self-Paced Kernel Estimation for Robust Blind Image Deblurring
Dong Gong, Mingkui Tan, Yanning Zhang, Anton van den Hengel, Qinfeng Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1661-1670

The challenge in blind image deblurring is to remove the effects of blur with limited prior information about the nature of the blur process. Existing methods often assume that the blur image is produced by linear convolution with additive Gaussian noise. However, including even a small number of outliers to this model in the kernel estimation process can significantly reduce the resulting image quality. Previous methods mainly rely on some simple but unreliable heuristics to identify outliers for kernel estimation. Rather than attempt to identify outliers to the model a priori, we instead propose to sequentially identify inliers, and gradually incorporate them into the estimation process. The self-paced kernel estimation scheme we propose represents a generalization of existing self-paced learning approaches, in which we gradually detect and include reliable inlier pixel sets in a blurred image for kernel estimation. Moreover, we automatically activate a subset of significant gradients w.r.t. the reliable inlier pixels, and then update the intermediate sharp image and the kernel accordingly. Experiments on both synthetic data and real-world images with various kinds of outliers demonstrate the effectiveness and robustness of the proposed method compared to the state-of-the-art methods.

****

Super-Trajectory for Video Segmentation
Wenguan Wang, Jianbing Shen, Jianwen Xie, Fatih Porikli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1671-1679

We introduce a novel semi-supervised video segmentation approach based on an efficient video representation, called as "super-trajectory". Each super-trajectory corresponds to a group of compact trajectories that exhibit consistent motion patterns, similar appearance and close spatiotemporal relationships. We generate trajectories using a probabilistic model, which handles occlusions and drifts in a robust and natural way. To reliably group trajectories, we adopt a modified version of the density peaks based clustering algorithm that allows capturing rich spatiotemporal relations among trajectories in the clustering process. The presented video representation is discriminative enough to accurately propagate the initial annotations in the first frame onto the remaining video frames. Extensive experimental analysis on challenging benchmarks demonstrate our method is capable of distinguishing the target objects from complex backgrounds and even reidentifying them after occlusions.

****

Be Your Own Prada: Fashion Synthesis With Structural Coherence
Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, Chen Change Loy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1680-1688

We present a novel and effective approach for generating new clothing on a wearer through generative adversarial learning. Given an input image of a person and a sentence describing a different outfit, our model "redresses" the person as desired, while at the same time keeping the wearer and her/his pose unchanged. Generating new outfits with precise regions conforming to a language description while retaining wearer's body structure is a new challenging task. Existing generative adversarial networks are not ideal in ensuring global coherence of structure given both the input photograph and language description as conditions. We address this challenge by decomposing the complex generative process into two conditional stages. In the first stage, we generate a plausible semantic segmentation map that obeys the wearer's pose as a latent spatial arrangement. An effective spatial constraint is formulated to guide the generation of this semantic segmentation map. In the second stage, a generative model with a newly proposed compositional mapping layer is used to render the final image with precise regions and textures conditioned on this map. We extended the DeepFashion dataset [8] by collecting sentence descriptions for 79K images. We demonstrate the effectiveness of our approach through both quantitative and qualitative evaluations. A user study is also conducted.

****

Wavelet-SRNet: A Wavelet-Based CNN for Multi-Scale Face Super Resolution
Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1689-1697

Most modern face super-resolution methods resort to convolutional neural networks (CNN) to infer high-resolution (HR) face images. When dealing with very low resolution (LR) images, the performance of these CNN based methods greatly degrades. Meanwhile, these methods tend to produce over-smoothed outputs and miss some textural details. To address these challenges, this paper presents a wavelet-bas ed CNN approach that can ultra-resolve a very low resolution face image of 16x16 or smaller pixel-size to its larger version of multiple scaling factors (2x, 4x, 8x and even 16x) in a unified framework. Different from conventional CNN methods directly inferring HR images, our approach firstly learns to predict the LR's corresponding series of HR's wavelet coefficients before reconstructing HR images from them. To capture both global topology information and local texture details of human faces, we present a flexible and extensible convolutional neural network with three types of loss: wavelet prediction loss, texture loss and full-image loss. Extensive experiments demonstrate that the proposed approach achieves more appealing results both quantitatively and qualitatively than state-of-the-art super-resolution methods.

****

Learning Gaze Transitions From Depth to Improve Video Saliency Estimation

George Leifman, Dmitry Rudoy, Tristan Swedish, Eduardo Bayro-Corrochano, Ramesh Raskar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1698-1707

In this paper we introduce a novel Depth-Aware Video Saliency approach to predict human focus of attention when viewing videos that contain a depth map (RGBD) on a 2D screen. Saliency estimation in this scenario is highly important since in the near future 3D video content will be easily acquired yet hard to display. Despite considerable progress in 3D display technologies, most are still expensive and require special glasses for viewing, so RGBD content is primarily viewed on 2D screens, removing the depth channel from the final viewing experience. We train a generative convolutional neural network that predicts the 2D viewing saliency map for a given frame using the RGBD pixel values and previous fixation estimates in the video. To evaluate the performance of our approach, we present a new comprehensive database of 2D viewing eye-fixation ground-truth for RGBD videos. Our experiments indicate that it is beneficial to integrate depth into video saliency estimates for content that is viewed on a 2D display. We demonstrate that our approach outperforms state-of-the-art methods for video saliency, achieving 15% relative improvement.

****

Joint Convolutional Analysis and Synthesis Sparse Representation for Single Image Layer Separation

Shuhang Gu, Deyu Meng, Wangmeng Zuo, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1708-1716

Analysis sparse representation (ASR) and synthesis sparse representation (SSR) are two representative approaches for sparsity-based image modeling. An image is described mainly by the non-zero coefficients in SSR, while it is characterized by the indices of zeros in ASR. To exploit the complementary representation mechanisms of ASR and SSR, we integrate the two models and propose a joint convolutional analysis and synthesis (JCAS) sparse representation model. The convolutional implementation is adopted to more effectively exploit the image global information. In JCAS, a single image is decomposed into two layers, one is approximated by ASR to represent image large-scale structures, and the other by SSR to represent image fine-scale textures. The synthesis dictionary is adaptively learned in JCAS to describe the texture patterns for different single image layer separation tasks. We evaluate the proposed JCAS model on a variety of applications, including rain streak removal, high dynamic range image tone mapping, etc. The results show that our JCAS method outperforms state-ofthe-arts in those applications in terms of both quantitative measure and visual perception quality.

****

Modelling the Scene Dependent Imaging in Cameras With a Deep Neural Network

Seonghyeon Nam, Seon Joo Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1717-1725

We present a novel deep learning framework that models the scene dependent image processing inside cameras. Often called as the radiometric calibration, the process of recovering RAW images from processed images (JPEG format in the sRGB color space) is essential for many computer vision tasks that rely on physically accurate radiance values. All previous works rely on the deterministic imaging mod el where the color transformation stays the same regardless of the scene and thus they can only be applied for images taken under the manual mode. In this paper, we propose a data-driven approach to learn the scene dependent and locally varying image processing inside cameras under the automode. Our method incorporates both the global and the local scene context into pixel-wise features via multi-scale pyramid of learnable histogram layers. The results show that we can model the imaging pipeline of different cameras that operate under the automode accurately in both directions (from RAW to sRGB, from sRGB to RAW) and we show how we can apply our method to improve the performance of image deblurring.

****

Transformed Low-Rank Model for Line Pattern Noise Removal

Yi Chang, Luxin Yan, Sheng Zhong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1726-1734

This paper addresses the problem of line pattern noise removal from a single image, such as rain streak, hyperspectral stripe and so on. Most of the previous methods model the line pattern noise in original image domain, which fail to explicitly exploit the directional characteristic, thus resulting in a redundant subspace with poor representation ability for those line pattern noise. To achieve a compact subspace for the line pattern structure, in this work, we incorporate a transformation into the image decomposition model so that maps the input image to a domain where the line pattern streak/stripe appearance has an extremely distinct low-rank structure, which naturally allows us to enforce a low-rank prior to extract the line pattern streak/stripe from the noisy image. Moreover, the random noise is usually mixed up with the line pattern noise, which makes the challenging problem much more difficult. While previous methods resort to the spectral or temporal correlation of the multi-images, we give a detailed analysis between the noisy and clean image in both local gradient and nonlocal domain, and propose a compositional directional total variational and low-rank prior for the image layer, thus to simultaneously accommodate both types of noise. The proposed method has been evaluated on two different tasks, including remote sensing image mixed random stripe noise removal and rain streak removal, all of which obtain very impressive performances.

****

Weakly Supervised Manifold Learning for Dense Semantic Object Correspondence

Utkarsh Gaur, B. S. Manjunath; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1735-1743

The goal of the semantic object correspondence problem is to compute dense association maps for a pair of images such that the same object parts get matched for very different appearing object instances. Our method builds on the recent findings that deep convolutional neural networks (DCNNs) implicitly learn a latent model of object parts even when trained for classification. We also leverage a key correspondence problem insight that the geometric structure between object parts is consistent across multiple object instances. These two concepts are then combined in the form of a novel optimization scheme. This optimization learns a feature embedding by rewarding for projecting features closer on the manifold if they have low feature-space distance. Simultaneously, the optimization penalizes feature clusters whose geometric structure is inconsistent with the observed geometric structure of object parts. In this manner, by accounting for feature space similarities and feature neighborhood context together, a manifold is learned where features belonging to semantically similar object parts cluster together. We also describe transferring these embedded features to the sister tasks of semantic keypoint classification and localization task via a Siamese DCNN. We provide qualitative results on the Pascal VOC 2012 images and quantitative results on the Pascal Berkeley dataset where we improve on the state of the art by over 5% on classification and over 9% on localization tasks.

****

Dual Motion GAN for Future-Flow Embedded Video Prediction

Xiaodan Liang, Lisa Lee, Wei Dai, Eric P. Xing; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1744-1752

Future frame prediction in videos is a promising avenue for unsupervised video r epresentation learning. Video frames are naturally generated by the inherent pixel flows from preceding frames based on the appearance and motion dynamics in the video. However, existing methods focus on directly hallucinating pixel values, resulting in blurry predictions. In this paper, we develop a dual motion Generative Adversarial Net (GAN) architecture, which learns to explicitly enforce future-frame predictions to be consistent with the pixel-wise flows in the video through a dual-learning mechanism. The primal future-frame prediction and dual future-flow prediction form a closed loop, generating informative feedback signals to each other for better video prediction. To make both synthesized future frames and flows indistinguishable from reality, a dual adversarial training method is proposed to ensure that the future-flow prediction is able to help infer realistic future-frames, while the future-frame prediction in turn leads to realistic optical flows. Our dual motion GAN also handles natural motion uncertainty in different pixel locations with a new probabilistic motion encoder, which is based on variational autoencoders. Extensive experiments demonstrate that the proposed dual motion GAN significantly outperforms state-of-the-art approaches on synthesizing new video frames and predicting future flows. Our model generalizes well across diverse visual scenes and shows superiority in unsupervised video representation learning.

****

Online Robust Image Alignment via Subspace Learning From Gradient Orientations
Qingqing Zheng, Yi Wang, Pheng-Ann Heng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1753-1762

Robust and efficient image alignment remains a challenging task, due to the massiveness of images, great illumination variations between images, partial occlusion and corruption. To address these challenges, we propose an online image alignment method via subspace learning from image gradient orientations (IGO). The proposed method integrates the subspace learning, transformed IGO reconstruction and image alignment into a unified online framework, which is robust for aligning images with severe intensity distortions. Our method is motivated by principal component analysis (PCA) from gradient orientations provides more reliable low-dimensional subspace than that from pixel intensities. Instead of processing in the intensity domain like conventional methods, we seek alignment in the IGO domain such that the aligned IGO of the newly arrived image can be decomposed as the sum of a sparse error and a linear composition of the IGO-PCA basis learned from previously well-aligned ones. The optimization problem is accomplished by an iterative linearization that minimizes the l1-norm of the sparse error. Furthermore, the IGO-PCA basis is adaptively updated based on incremental thin singular value decomposition (SVD) which takes the shift of IGO mean into consideration. The efficacy of the proposed method is validated on extensive challenging datasets through image alignment and face recognition. Experimental results demonstrate that our algorithm provides more illumination- and occlusion-robust image alignment than state-of-the-art methods do.

****

Learning Dynamic Siamese Network for Visual Object Tracking
Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, Song Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1763-1771

How to effectively learn temporal variation of target appearance, to exclude the interference of cluttered background, while maintaining real-time response, is an essential problem of visual object tracking. Recently, Siamese networks have shown great potentials of matching based trackers in achieving balanced accuracy and beyond real-time speed. However, they still have a big gap to classification & updating based trackers in tolerating the temporal changes of objects and imaging conditions. In this paper, we propose dynamic Siamese network, via a fast transformation learning model that enables effective online learning of target appearance variation and background suppression from previous frames. We then present elementwise multi-layer fusion to adaptively integrate the network outputs using multi-level deep features. Unlike state-of-the-art trackers, our approach allows the usage of any feasible generally- or particularly-trained features, such as SiamFC and VGG. More importantly, the proposed dynamic Siamese network can be jointly trained as a whole directly on the labeled video sequences, thus can take full advantage of the rich spatial temporal information of moving objects. As a result, our approach achieves state-of-the-art performance on OTB-2013 and VOT-2015 benchmarks, while exhibits superiorly balanced accuracy and real-time response over state-of-the-art competitors.

****

High Order Tensor Formulation for Convolutional Sparse Coding

Adel Bibi, Bernard Ghanem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1772-1780

Convolutional sparse coding (CSC) has gained attention for its successful role as a reconstruction and a classification tool in the computer vision and machine learning community. Current CSC methods can only reconstruct single-feature 2D images independently. However, learning multi-dimensional dictionaries and sparse codes for the reconstruction of multi-dimensional data is very important, as it examines correlations among all the data jointly. This provides more capacity for the learned dictionaries to better reconstruct data. In this paper, we propose a generic and novel formulation for the CSC problem that can handle an arbitrary order tensor of data. Backed with experimental results, our proposed formulation can not only tackle applications that are not possible with standard CSC solvers, including colored video reconstruction (5D- tensors), but it also performs favorably in reconstruction with much fewer parameters as compared to naive extensions of standard CSC to multiple features/channels.

****

Learning Proximal Operators: Using Denoising Networks for Regularizing Inverse Imaging Problems

Tim Meinhardt, Michael Moller, Caner Hazirbas, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1781-1790

While variational methods have been among the most powerful tools for solving linear inverse problems in imaging, deep (convolutional) neural networks have recently taken the lead in many challenging benchmarks. A remaining drawback of deep learning approaches is their requirement for an expensive retraining whenever the specific problem, the noise level, noise type, or desired measure of fidelity changes. On the contrary, variational methods have a plug-and-play nature as they usually consist of separate data fidelity and regularization terms. In this paper we study the possibility of replacing the proximal operator of the regularization used in many convex energy minimization algorithms by a denoising neural network. The latter therefore serves as an implicit natural image prior, while the data term can still be chosen independently. Using a fixed denoising neural network in exemplary problems of image deconvolution with different blur kernels and image demosaicking, we obtain state-of-the-art reconstruction results. These indicate the high generalizability of our approach and a reduction of the need for problem-specific training. Additionally, we discuss novel results on the analysis of possible optimization algorithms to incorporate the network into, as well as the choices of algorithm parameters and their relation to the noise level the neural network is trained on.

****

ScaleNet: Guiding Object Proposal Generation in Supermarkets and Beyond

Siyuan Qiao, Wei Shen, Weichao Qiu, Chenxi Liu, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1791-1800

Motivated by product detection in supermarkets, this paper studies the problem of object proposal generation in supermarket images and other natural images. We argue that estimation of object scales in images is helpful for generating object proposals, especially for supermarket images where object scales are usually within a small range. Therefore, we propose to estimate object scales of images before generating object proposals. The proposed method for predicting object scales is called ScaleNet. To validate the effectiveness of ScaleNet, we build three supermarket datasets, two of which are real-world datasets used for testing and the other one is a synthetic dataset used for training. In short, we extend the previous state-of-the-art object proposal methods by adding a scale prediction phase. The resulted method outperforms the previous state-of-the-art on the sup ermarket datasets by a large margin. We also show that the approach works for object proposal on other natural images and it outperforms the previous state-of-the-art object proposal methods on the MS COCO dataset. The supermarket datasets, the virtual supermarkets, and the tools for creating more synthetic datasets will be made public.

****

Temporal Dynamic Graph LSTM for Action-Driven Video Object Detection

Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1801-1810

In this paper, we investigate a weakly-supervised object detection framework. Most existing frameworks focus on using static images to learn object detectors. However, these detectors often fail to generalize to videos because of the existing domain shift. Therefore, we investigate learning these detectors directly from boring videos of daily activities. Instead of using bounding boxes, we explore the use of action descriptions as supervision since they are relatively easy to gather. A common issue, however, is that objects of interest that are not involved in human actions are often absent in global action descriptions known as "missing label". To tackle this problem, we propose a novel temporal dynamic graph Long Short-Term Memory network (TD- Graph LSTM). TD-Graph LSTM enables global temporal reasoning by constructing a dynamic graph that is based on temporal correlations of object proposals and spans the entire video. The missing label issue for each individual frame can thus be significantly alleviated by transferring knowledge across correlated objects proposals in the whole video. Extensive evaluations on a large-scale daily-life action dataset (i.e., Charades) demonstrates the superiority of our proposed method. We also release object bounding-box annotations for more than 5,000 frames in Charades. We believe this annotated data can also benefit other research on video-based object recognition in the future.

****

VQS: Linking Segmentations to Questions and Answers for Supervised Attention in VQA and Question-Focused Semantic Segmentation

Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, Boqing Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1811-1820

Rich and dense human labeled datasets are the main enabling factor, among others , for the recent exciting work on vision-language understanding. Many seemingly distinct annotations (e.g., semantic segmentation and visual questions answering (VQA)) are inherently connected in that they reveal different levels and perspectives of human understandings about the same visual scenes --- and even the same set of MS COCO images. The popularity of MS COCO could strongly correlate those annotations and tasks. Explicitly linking them up, as we envision, can significantly benefit not only individual tasks but also the overarching goal of unified vision-language understand. We present the preliminary work of linking the instance segmentations provided by MS COCO to the questions and answers (QA) in the VQA dataset. We call the collected links visual questions and segmentation answers (VQS). They transfer human supervision between the previously separate tasks , offer more effective leverage to existing problems, and also open the door for new tasks and richer models. We study two applications of the VQS data in this paper: supervised attention for VQA and a novel question-focused semantic segmentation task. For the former, we obtain state-of-the-art results on the VQA real multiple-choice task by simply augmenting multilayer perceptrons with some attention features that are learned by using the segmentation-QA links as explicit supervision. To put the latter in perspective, we study two plausible methods and an oracle upper bound.

****

Multi-Modal Factorized Bilinear Pooling With Co-Attention Learning for Visual Question Answering

Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1821-1830

Visual question answering (VQA) is challenging because it requires a simultaneous understanding of both the visual content of images and the textual content of

questions. The approaches used to represent the images and questions in a fine-grained manner and questions and to fuse these multi-modal features play key roles in performance. Bilinear pooling based models have been shown to outperform traditional linear models for VQA, but their high-dimensional representations and high computational complexity may seriously limit their applicability in practice. For multi-modal feature fusion, here we develop a Multi-modal Factorized Bilinear (MFB) pooling approach to efficiently and effectively combine multi-modal features, which results in superior performance for VQA compared with other bilinear pooling approaches. For fine-grained image and question representation, we develop a co-attention mechanism using an end-to-end deep network architecture to jointly learn both the image and question attentions. Combining the proposed MFB approach with co-attention learning in a new network architecture provides a unified model for VQA. Our experimental results demonstrate that the single MFB with co-attention model achieves new state-of-the-art performance on the real-world VQA dataset. Code available at https://github.com/yuzcccc/mfb
****
SCNet: Learning Semantic Correspondence
Kai Han, Rafael S. Rezende, Bumsub Ham, Kwan-Yee K. Wong, Minsu Cho, Cordelia Schmid, Jean Ponce; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1831-1840
This paper addresses the problem of establishing semantic correspondences between images depicting different instances of the same object or scene category. Previous approaches focus on either combining a spatial regularizer with hand-crafted features, or learning a correspondence model for appearance only. We propose instead a convolutional neural network architecture, called SCNet, for learning a geometrically plausible model for semantic correspondence. SCNet uses region proposals as matching primitives, and explicitly incorporates geometric consistency in its loss function. It is trained on image pairs obtained from the PASCAL VOC 2007 keypoint dataset, and a comparative evaluation on several standard benchmarks demonstrates that the proposed approach substantially outperforms both recent deep learning architectures and previous methods based on hand-crafted features.
****
Soft Proposal Networks for Weakly Supervised Object Localization
Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, Jianbin Jiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1841-1850
Weakly supervised object localization remains challenging, where only image labels instead of bounding boxes are available during training. Object proposal is an effective component in localization, but often computationally expensive and incapable of joint optimization with some of the remaining modules. In this paper, to the best of our knowledge, we for the first time integrate weakly supervised object proposal into convolutional neural networks (CNNs) in an end-to-end learning manner. We design a network component, Soft Proposal (SP), to be plugged into any standard convolutional architecture to introduce the nearly cost-free object proposal, orders of magnitude faster than state-of-the-art methods. In the SP-augmented CNNs, referred to as Soft Proposal Networks (SPNs), iteratively evolved object proposals are generated based on the deep feature maps then projected back, and further jointly optimized with network parameters, with image-level supervision only. Through the unified learning process, SPNs learn better object-centric filters, discover more discriminative visual evidence, and suppress background interference, significantly boosting both weakly supervised object localization and classification performance. We report the best results on popular benchmarks, including PASCAL VOC, MS COCO, and ImageNet.
****
Class Rectification Hard Mining for Imbalanced Deep Learning
Qi Dong, Shaogang Gong, Xiatian Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1851-1860
Recognising detailed facial or clothing attributes in images of people is a challenging task for computer vision, especially when the training data are both in very large scale and extremely imbalanced among different attribute classes. To

address this problem, we formulate a novel scheme for batch incremental hard sample mining of minority attribute classes from imbalanced large scale training data. We develop an end-to-end deep learning framework capable of avoiding the dominant effect of majority classes by discovering sparsely sampled boundaries of minority classes. This is made possible by introducing a Class Rectification Loss (CRL) regularising algorithm. We demonstrate the advantages and scalability of CRL over existing state-of-the-art attribute recognition and imbalanced data learning models on two large scale imbalanced benchmark datasets, the CelebA facial attribute dataset and the X-Domain clothing attribute dataset.

****

Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs
Vishwanath A. Sindagi, Vishal M. Patel; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1861-1870

We present a novel method called Contextual Pyramid CNN (CP-CNN) for generating high-quality crowd density and count estimation by explicitly incorporating global and local contextual information of crowd images. The proposed CP-CNN consists of four modules: Global Context Estimator (GCE), Local Context Estimator (LCE), Density Map Estimator (DME) and a Fusion-CNN (F-CNN). GCE is a VGG-16 based CNN that encodes global context and it is trained to classify input images into different density classes, whereas LCE is another CNN that encodes local context information and it is trained to perform patch-wise classification of input images into different density classes. DME is a multi-column architecture-based CNN that aims to generate high-dimensional feature maps from the input image which are fused with the contextual information estimated by GCE and LCE using F-CNN. To generate high resolution and high-quality density maps, F-CNN uses a set of convolutional and fractionally-strided convolutional layers and it is trained along with the DME in an end-to-end fashion using a combination of adversarial loss and pixel-level Euclidean loss. Extensive experiments on highly challenging datasets show that the proposed method achieves significant improvements over the state-of-the-art methods.

****

See the Glass Half Full: Reasoning About Liquid Containers, Their Volume and Content
Roozbeh Mottaghi, Connor Schenck, Dieter Fox, Ali Farhadi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1871-1880

Humans have rich understanding of liquid containers and their contents; for example, we can effortlessly pour water from a pitcher to a cup. Doing so requires estimating the volume of the cup, approximating the amount of water in the pitcher, and predicting the behavior of water when we tilt the pitcher. Very little attention in computer vision has been made to liquids and their containers. In this paper, we study liquid containers and their contents, and propose methods to estimate the volume of containers, approximate the amount of liquid in them, and perform comparative volume estimations all from a single RGB image. Furthermore, we show the results of the proposed model for predicting the behavior of liquids inside containers when one tilts the containers. We also introduce a new dataset of Containers Of liQuid contEnt (COQE) that contains more than 5,000 images of 10,000 liquid containers in context labelled with volume, amount of content, bounding box annotation, and corresponding similar 3D CAD models.

****

Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding
Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, Gang Hua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1881-1889

We address the problem of dense visual-semantic embedding that maps not only full sentences and whole images but also phrases within sentences and salient regions within images into a multimodal embedding space. As a result, we can produce several region-oriented and expressive phrases rather than just an overview sentence to describe an image. In particular, we present a hierarchical structured recurrent neural network (RNN), namely Hierarchical Multimodal LSTM (HM-LSTM) model. Different from chain structured RNN, our model presents a hierarchical structure so that it can naturally build representations for phrases and image region s, and further exploit their hierarchical relations. Moreover, the fine-grained correspondences between phrases and image regions can be automatically learned and utilized to boost the learning of the dense embedding space. Extensive experiments on several datasets validate the efficacy of our proposed method, which compares favorably with the state-of-the-art methods.

****

Identity-Aware Textual-Visual Matching With Latent Co-Attention

Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1890-1899

Textual-visual matching aims at measuring similarities between sentence descriptions and images. Most existing methods tackle this problem without effectively utilizing identity-level annotations. In this paper, we propose an identity-aware two-stage framework for the textual-visual matching problem. Our stage-1 CNN-LSTM network learns to embed cross-modal features with a novel Cross-Modal Cross-Entropy (CMCE) loss. The stage-1 network is able to efficiently screen easy incorrect matchings and also provide initial training point for the stage-2 training. The stage-2 CNN-LSTM network refines the matching results with a latent co-attention mechanism. The spatial attention relates each word with corresponding image regions while the latent semantic attention aligns different sentence structures to make the matching results more robust to sentence structure variations. Extensive experiments on three datasets with identity-level annotations show that our framework outperforms state-of-the-art approaches by large margins.

****

Learning Deep Neural Networks for Vehicle Re-ID With Visual-Spatio-Temporal Path Proposals

Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1900-1909

Vehicle re-identification is an important problem and has many applications in video surveillance and intelligent transportation. It gains increasing attention because of the recent advances of person re-identification techniques. However, unlike person re-identification, the visual differences between pairs of vehicle images are usually subtle and even challenging for humans to distinguish. Incorporating additional spatio-temporal information is vital for solving the challenging re-identification task. Existing vehicle re-identification methods ignored or used over-simplified models for the spatio-temporal relations between vehicle images. In this paper, we propose a two-stage framework that incorporates complex spatio-temporal information for effectively regularizing the re-identification results. Given a pair of vehicle images with their spatio-temporal information, a candidate visual-spatio-temporal path is first generated by a chain MRF model with a deeply learned potential function, where each visual-spatio-temporal state corresponds to an actual vehicle image with its spatio-temporal information. A Siamese-CNN+Path-LSTM model takes the candidate path as well as the pairwise queries to generate their similarity score. Extensive experiments and analysis show the effectiveness of our proposed method and individual components.

****

Learning From Noisy Labels With Distillation

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, Li-Jia Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1910-1918

The ability of learning from noisy labels is very useful in many visual recognition tasks, as a vast amount of data with noisy labels are relatively easy to obtain. Traditionally, label noise has been treated as statistical outliers, and techniques such as importance re-weighting and bootstrapping have been proposed to alleviate the problem. According to our observation, the real-world noisy labels exhibit multi-mode characteristics as the true labels, rather than behaving like independent random outliers. In this work, we propose a unified distillation framework to use "side" information, including a small clean dataset and label relations in knowledge graph, to "hedge the risk" of learning from noisy labels. Unlike the traditional approaches evaluated based on simulated label noises, we propose a suite of new benchmark datasets, in Sports, Species and Artifacts doma ins, to evaluate the task of learning from noisy labels in the practical setting. The empirical study demonstrates the effectiveness of our proposed method in all the domains.

****

DSOD: Learning Deeply Supervised Object Detectors From Scratch

Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, Xiangyang Xue; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1919-1927

We present Deeply Supervised Object Detector (DSOD), a framework that can learn object detectors from scratch. State-of-the-art object objectors rely heavily on the off-the-shelf networks pre-trained on large-scale classification datasets like ImageNet, which incurs learning bias due to the difference on both the loss functions and the category distributions between classification and detection tasks. Model fine-tuning for the detection task could alleviate this bias to some extent but not fundamentally. Besides, transferring pre-trained models from classification to detection between discrepant domains is even more difficult (e.g. RGB to depth images). A better solution to tackle these two critical problems is to train object detectors from scratch, which motivates our proposed DSOD. Previous efforts in this direction mostly failed due to much more complicated loss functions and limited training data in object detection. In DSOD, we contribute a set of design principles for training object detectors from scratch. One of the key findings is that deep supervision, enabled by dense layer-wise connections, plays a critical role in learning a good detector. Combining with several other principles, we develop DSOD following the single-shot detection (SSD) framework. Experiments on PASCAL VOC 2007, 2012 and MS COCO datasets demonstrate that DSOD can achieve better results than the state-of-the-art solutions with much more compact models. For instance, DSOD outperforms SSD on all three benchmarks with real-time detection speed, while requires only 1/2 parameters to SSD and 1/10 parameters to Faster RCNN.

****

Phrase Localization and Visual Relationship Detection With Comprehensive Image-Language Cues

Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, Svetlana Lazebnik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1928-1937

This paper presents a framework for localization or grounding of phrases in images using a large collection of linguistic and visual cues. We model the appearance, size, and position of entity bounding boxes, adjectives that contain attribute information, and spatial relationships between pairs of entities connected by verbs or prepositions. Special attention is given to relationships between people and clothing or body part mentions, as they are useful for distinguishing individuals. We automatically learn weights for combining these cues and at test time, perform joint inference over all phrases in a caption. The resulting system produces state of the art performance on phrase localization on the Flickr30k Entities dataset and visual relationship detection on the Stanford VRD dataset.

****

Chained Cascade Network for Object Detection

Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1938-1946

Cascade is a widely used approach that rejects obvious negative samples at early stages for learning better classifier and faster inference. This paper presents chained cascade network (CC-Net). In this CC-Net, there are many cascade stages. Preceding cascade stages are placed at shallow layers. Easy hard examples are rejected at shallow layers so that the computation for deeper or wider layers is not required. In this way, features and classifiers at latter stages handle more difficult samples with the help of features and classifiers in previous stages. It yields consistent boost in detection performance on PASCAL VOC 2007 and ImageNet for both fast RCNN and Faster RCNN. CC-Net saves computation for both training and testing. Code is available on.

****

VPGNet: Vanishing Point Guided Network for Lane and Road Marking Detection and Recognition

Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1947-1955

In this paper, we propose a unified end-to-end trainable multi-task network that jointly handles lane and road marking detection and recognition that is guided by a vanishing point under adverse weather conditions. We tackle rainy and low illumination conditions, which have not been extensively studied until now due to clear challenges. For example, images taken under rainy days are subject to low illumination, while wet roads cause light reflection and distort the appearance of lane and road markings. At night, color distortion occurs under limited illumination. As a result, no benchmark dataset exists and only a few developed algorithms work under poor weather conditions. To address this shortcoming, we build up a lane and road marking benchmark which consists of about 20,000 images with 17 lane and road marking classes under four different scenarios: no rain, rain, heavy rain, and night. We train and evaluate several versions of the proposed multi-task network and validate the importance of each task. The resulting approach, VPGNet, can detect and classify lanes and road markings, and predict a vanishing point with a single forward pass. Experimental results show that our approach achieves high accuracy and robustness under various conditions in real-time (20 fps). The benchmark and the VPGNet model will be publicly available.

****

Unsupervised Learning of Important Objects From First-Person Videos

Gedas Bertasius, Hyun Soo Park, Stella X. Yu, Jianbo Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1956-1964

A first-person camera, placed at a person's head, captures, which objects are important to the camera wearer. Most prior methods for this task learn to detect such important objects from the manually labeled first-person data in a supervised fashion. However, important objects are strongly related to the camera wearer's internal state such as his intentions and attention, and thus, only the person wearing the camera can provide the importance labels. Such a constraint makes the annotation process costly and limited in scalability. In this work, we show that we can detect important objects in first-person images without the supervision by the camera wearer or even third-person labelers. We formulate an important detection problem as an interplay between the 1) segmentation and 2) recognition agents. The segmentation agent first proposes a possible important object segmentation mask for each image, and then feeds it to the recognition agent, which learns to predict an important object mask using visual semantics and spatial features. We implement such an interplay between both agents via an alternating cross-pathway supervision scheme inside our proposed Visual-Spatial Network (VSN). Our VSN consists of spatial ("where") and visual ("what") pathways, one of which learns common visual semantics while the other focuses on the spatial location cues. Our unsupervised learning is accomplished via a cross-pathway supervision, where one pathway feeds its predictions to a segmentation agent, which proposes a candidate important object segmentation mask that is then used by the other pathway as a supervisory signal. We show our method's success on two different important object datasets, where our method achieves similar or better results as the supervised methods.

****

An Analysis of Visual Question Answering Algorithms

Kushal Kafle, Christopher Kanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1965-1973

In visual question answering (VQA), an algorithm must answer text-based questions about images. While multiple datasets for VQA have been created since late 2014, they all have flaws in both their content and the way algorithms are evaluated on them. As a result, evaluation scores are inflated and predominantly determined by answering easier questions, making it difficult to compare different methods. In this paper, we analyze existing VQA algorithms using a new dataset called the Task Driven Image Understanding Challenge (TDIUC), which has over 1.6 mill ion questions organized into 12 different categories. We also introduce questions that are meaningless for a given image to force a VQA system to reason about image content. We propose new evaluation schemes that compensate for over-represented question-types and make it easier to study the strengths and weaknesses of algorithms. We analyze the performance of both baseline and state-of-the-art VQA models, including multi-modal compact bilinear pooling (MCB), neural module networks, and recurrent answering units. Our experiments establish how attention helps certain categories more than others, determine which models work better than others, and explain how simple models (e.g. MLP) can surpass more complex models (MCB) by simply learning to answer large, easy question categories.

****

Visual Relationship Detection With Internal and External Linguistic Knowledge Distillation

Ruichi Yu, Ang Li, Vlad I. Morariu, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1974-1982

Understanding the visual relationship between two objects involves identifying the subject, the object, and a predicate relating them.We leverage the strong correlations between the predicate and the (subj,obj) pair (both semantically and spatially) to predict predicates conditioned on the subjects and the objects. Modeling the three entities jointly more accurately reflects their relationships compared to modeling them independently, but it complicates learning since the semantic space of visual relationships is huge and training data is limited, especially for long-tail relationships that have few instances. To overcome this, we use knowledge of linguistic statistics to regularize visual model learning. We obtain linguistic knowledge by mining from both training annotations (internal knowledge) and publicly available text, e.g., Wikipedia (external knowledge), computing the conditional probability distribution of a predicate given a (subj,obj) pair. As we train the visual model, we distill this knowledge into the deep model to achieve better generalization. Our experimental results on the Visual Relationship Detection (VRD) and Visual Genome datasets suggest that with this linguistic knowledge distillation, our model outperforms the state-of-the-art methods significantly, especially when predicting unseen relationships (e.g., recall improved from 8.45% to 19.17% on VRD zero-shot testing set).

****

A Two Stream Siamese Convolutional Neural Network for Person Re-Identification

Dahjung Chung, Khalid Tahboub, Edward J. Delp; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1983-1991

Person re-identification is an important task in video surveillance systems. It can be formally defined as establishing the correspondence between images of a person taken from different cameras at different times. In this pa- per, we present a two stream convolutional neural network where each stream is a Siamese network. This architecture can learn spatial and temporal information separately. We also propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective of predicting a person's identity. We evaluate our proposed method on the publicly available PRID2011 and iLIDS-VID datasets and demonstrate the efficacy of our proposed method. On average, the top rank matching accuracy is 4% higher than the accuracy achieved by the cross-view quadratic discriminant analysis used in combination with the hierarchical Gaussian descriptor (GOG+XQDA), and 5% higher than the recurrent neural network method.

****

No More Discrimination: Cross City Adaptation of Road Scene Segmenters

Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, Min Sun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1992-2001

Despite the recent success of deep-learning based semantic segmentation, deploying a pre-trained road scene segmenter to a city whose images are not presented in the training set would not achieve satisfactory performance due to dataset biases. Instead of collecting a large number of annotated images of each city of interest to train or refine the segmenter, we propose an unsupervised learning app

roach to adapt road scene segmenters across different cities. By utilizing Google Street View and its time-machine feature, we can collect unannotated images for each road scene at different times, so that the associated static-object priors can be extracted accordingly. By advancing a joint global and class-specific domain adversarial learning framework, adaptation of pre-trained segmenters to that city can be achieved without the need of any user annotation or interaction. We show that our method improves the performance of semantic segmentation in multiple cities across continents, while it performs favorably against state-of-the-art approaches requiring annotated training data.
****

Open Vocabulary Scene Parsing
Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, Antonio Torralba; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2002-2010
Recognizing arbitrary objects in the wild has been a challenging problem due to the limitations of existing classification models and datasets. In this paper, we propose a new task that aims at parsing scenes with a large and open vocabulary, and several evaluation metrics are explored for this problem. Our approach is a joint image pixel and word concept embeddings framework, where word concepts are connected by semantic relations. We validate the open vocabulary prediction ability of our framework on ADE20K dataset which covers a wide variety of scenes and objects. We further explore the trained joint embedding space to show its interpretability.
****

Learned Watershed: End-To-End Learning of Seeded Segmentation
Steffen Wolf, Lukas Schott, Ullrich Kothe, Fred Hamprecht; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2011-2019
Learned boundary maps are known to outperform hand-crafted ones as a basis for the watershed algorithm. We show, for the first time, how to train watershed computation jointly with boundary map prediction. The estimator for the merging priorities is cast as a neural network that is convolutional (over space) and recurrent (over iterations). The latter allows learning of complex shape priors. The method gives the best known seeded segmentation results on the CREMI segmentation challenge.
****

Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes
Yang Zhang, Philip David, Boqing Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2020-2030
During the last half decade, convolutional neural networks (CNNs) have triumphed over semantic segmentation, which is a core task of various emerging industrial applications such as autonomous driving and medical imaging. However, to train CNNs requires a huge amount of data, which is difficult to collect and laborious to annotate. Recent advances in computer graphics make it possible to train CNN models on photo-realistic synthetic data with computer-generated annotations. Despite this, the domain mismatch between the real images and the synthetic data significantly decreases the models' performance. Hence we propose a curriculum-style learning approach to minimize the domain gap in semantic segmentation. The curriculum domain adaptation solves easy tasks first in order to infer some necessary properties about the target domain; in particular, the first task is to learn global label distributions over images and local distributions over landmark superpixels. These are easy to estimate because images of urban traffic scenes have strong idiosyncrasies (e.g., the size and spatial relations of buildings, streets, cars, etc.). We then train the segmentation network in such a way that the network predictions in the target domain follow those inferred properties. In experiments, our method significantly outperforms the baselines as well as the only known existing approach to the same problem.
****

Scale-Adaptive Convolutions for Scene Parsing
Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2031-2039

Many existing scene parsing methods adopt Convolutional Neural Networks with fixed-size receptive fields, which frequently result in inconsistent predictions of large objects and invisibility of small objects. To tackle this issue, we propose a scale-adaptive convolution to acquire flexible-size receptive fields during scene parsing. Through adding a new scale regression layer, we can dynamically infer the position-adaptive scale coefficients which are adopted to resize the convolutional patches. Consequently, the receptive fields can be adjusted automatically according to the various sizes of the objects in scene images. Thus, the problems of invisible small objects and inconsistent large-object predictions can be alleviated. Furthermore, our proposed scale-adaptive convolutions are not only differentiable to learn the convolutional parameters and scale coefficients in an end-to-end way, but also of high parallelizability for the convenience of GPU implementation. Additionally, since the new scale regression layers are learned implicitly, any extra training supervision of object sizes is unnecessary. Extensive experiments on Cityscapes and ADE20K datasets well demonstrate the effectiveness of the proposed scale-adaptive convolutions.

****

Privacy-Preserving Visual Learning Using Doubly Permuted Homomorphic Encryption
Ryo Yonetani, Vishnu Naresh Boddeti, Kris M. Kitani, Yoichi Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2040-2050

We propose a privacy-preserving framework for learning visual classifiers by leveraging distributed private image data. This framework is designed to aggregate multiple classifiers updated locally using private data and to ensure that no private information about the data is exposed during and after its learning procedure. We utilize a homomorphic cryptosystem that can aggregate the local classifiers while they are encrypted and thus kept secret. To overcome the high computational cost of homomorphic encryption of high-dimensional classifiers, we (1) impose sparsity constraints on local classifier updates and (2) propose a novel efficient encryption scheme named doubly-permuted homomorphic encryption (DPHE) which is tailored to sparse high-dimensional data. DPHE (i) decomposes sparse data into its constituent non-zero values and their corresponding support indices, (ii) applies homomorphic encryption only to the non-zero values, and (iii) employs double permutations on the support indices to make them secret. Our experimental evaluation on several public datasets shows that the proposed approach achieves comparable performance against state-of-the-art visual recognition methods while preserving privacy and significantly outperforms other privacy-preserving methods.

****

Multi-Task Self-Supervised Visual Learning
Carl Doersch, Andrew Zisserman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2051-2060

We investigate methods for combining multiple self-supervised tasks---i.e., supervised tasks where data can be collected without manual labeling---in order to train a single visual representation. First, we provide an apples-to-apples comparison of four different self-supervised tasks using the very deep ResNet-101 architecture. We then combine tasks to jointly train a network. We also explore lasso regularization to encourage the network to factorize the information in its representation, and methods for "harmonizing" network inputs in order to learn a more unified representation. We evaluate all methods on ImageNet classification, PASCAL VOC detection, and NYU depth prediction. Our results show that deeper networks work better, and that combining tasks---even via a naive multi-head architecture---always improves performance. Our best joint network nearly matches the PASCAL performance of a model pre-trained on ImageNet classification, and matches the ImageNet network on NYU depth prediction.

****

A Self-Balanced Min-Cut Algorithm for Image Clustering
Xiaojun Chen, Joshua Zhexue Haung, Feiping Nie, Renjie Chen, Qingyao Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2061-2069

Many spectral clustering algorithms have been proposed and successfully applied to image data analysis such as content based image retrieval, image annotation, and image indexing. Conventional spectral clustering algorithms usually involve a two-stage process: eigendecomposition of similarity matrix and clustering assignments from eigenvectors by k-means or spectral rotation. However, the final clustering assignments obtained by the two-stage process may deviate from the assignments by directly optimize the original objective function. Moreover, most of these methods usually have very high computational complexities. In this paper, we propose a new min-cut algorithm for image clustering, which scales linearly to the data size. In the new method, a self-balanced min-cut model is proposed in which the Exclusive Lasso is implicitly introduced as a balance regularizer in order to produce balanced partition. We propose an iterative algorithm to solve the new model, which has a time complexity of O(n) where n is the number of samples. Theoretical analysis reveals that the new method can simultaneously minimize the graph cut and balance the partition across all clusters. A series of experiments were conducted on both synthetic and benchmark data sets and the experimental results show the superior performance of the new method.

****

Is Second-Order Information Helpful for Large-Scale Visual Recognition?

Peihua Li, Jiangtao Xie, Qilong Wang, Wangmeng Zuo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2070-2078

By stacking layers of convolution and nonlinearity, convolutional networks (ConvNets) effectively learn from low-level to high-level features and discriminative representations. Since the end goal of large-scale recognition is to delineate complex boundaries of thousands of classes, adequate exploration of feature distributions is important for realizing full potentials of ConvNets. However, state-of-the-art works concentrate only on deeper or wider architecture design, while rarely exploring feature statistics higher than first-order. We take a step towards addressing this problem. Our method consists in covariance pooling, instead of the most commonly used first-order pooling, of high-level convolutional features. The main challenges involved are robust covariance estimation given a small sample of large-dimensional features and usage of the manifold structure of covariance matrices. To address these challenges, we present a Matrix Power Normalized Covariance (MPN-COV) method. We develop forward and backward propagation formulas regarding the nonlinear matrix functions such that MPN-COV can be trained end-to-end. In addition, we analyze both qualitatively and quantitatively its advantage over the well-known Log-Euclidean metric. On the ImageNet 2012 validation set, by combining MPN-COV we achieve over 4%, 3% and 2.5% gains for AlexNet, VGG-M and VGG-16, respectively; integration of MPN-COV into 50-layer ResNet outperforms ResNet-101 and is comparable to ResNet-152. The source code will be available on the project page: http://www.peihuali.org/MPN-COV.

****

Factorized Bilinear Models for Image Recognition

Yanghao Li, Naiyan Wang, Jiaying Liu, Xiaodi Hou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2079-2087

Although Deep Convolutional Neural Networks (CNNs) have liberated their power in various computer vision tasks, the most important components of CNN, convolutional layers and fully connected layers, are still limited to linear transformations. In this paper, we propose a novel Factorized Bilinear (FB) layer to model the pairwise feature interactions by considering the quadratic terms in the transformations. Compared with existing methods that tried to incorporate complex non-linearity structures into CNNs, the factorized parameterization makes our FB layer only require a linear increase of parameters and affordable computational cost. To further reduce the risk of overfitting of the FB layer, a specific remedy called DropFactor is devised during the training process. We also analyze the connection between FB layer and some existing models, and show FB layer is a generalization to them. Finally, we validate the effectiveness of FB layer on several widely adopted datasets including CIFAR-10, CIFAR-100 and ImageNet, and demonstrate superior results compared with various state-of-the-art deep models.

****

Octree Generating Networks: Efficient Convolutional Architectures for High-Resolution 3D Outputs

Maxim Tatarchenko, Alexey Dosovitskiy, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2088-2096

We present a deep convolutional decoder architecture that can generate volumetric 3D outputs in a compute- and memory-efficient manner by using an octree representation. The network learns to predict both the structure of the octree, and the occupancy values of individual cells. This makes it a particularly valuable technique for generating 3D shapes. In contrast to standard decoders acting on regular voxel grids, the architecture does not have cubic complexity. This allows representing much higher resolution outputs with a limited memory budget. We demonstrate this in several application domains, including 3D convolutional autoencoders, generation of objects and whole scenes from high-level representations, and shape from a single image.

****

Truncating Wide Networks Using Binary Tree Architectures

Yan Zhang, Mete Ozay, Shuohao Li, Takayuki Okatani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2097-2105

In this paper, we propose a binary tree architecture to truncate architecture of wide networks by reducing the width of the networks. More precisely, in the proposed architecture, the width is incrementally reduced from lower layers to higher layers in order to increase the expressive capacity of networks with a less increase on parameter size. Also, in order to ease the gradient vanishing problem, features obtained at different layers are concatenated to form the output of our architecture. By employing the proposed architecture on a baseline wide network, we can construct and train a new network with same depth but considerably less number of parameters. In our experimental analyses, we observe that the proposed architecture enables us to obtain better parameter size and accuracy trade-off compared to baseline networks using various benchmark image classification datasets. The results show that our model can decrease the classification error of a baseline from 20.43% to 19.22% on Cifar-100 using only 28% of parameters that the baseline has.

****

Bringing Background Into the Foreground: Making All Classes Equal in Weakly-Supervised Video Semantic Segmentation

Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, Jose M. Alvarez; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2106-2116

Pixel-level annotations are expensive and time-consuming to obtain. Hence, weak supervision using only image tags could have a significant impact in semantic segmentation. Recent years have seen great progress in weakly-supervised semantic segmentation, whether from a single image or from videos. However, most existing methods are designed to handle a single background class. In practical applications, such as autonomous navigation, it is often crucial to reason about multiple background classes. In this paper, we introduce an approach to doing so by making use of classifier heatmaps. We then develop a two-stream deep architecture that jointly leverages appearance and motion, and design a loss based on our heatmaps to train it. Our experiments demonstrate the benefits of our classifier heatmaps and of our two-stream architecture on challenging urban scene datasets and on the YouTube-Objects benchmark, where we obtain state-of-the-art results.

****

View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition From Skeleton Data

Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, Nanning Zheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2117-2126

Skeleton-based human action recognition has recently attracted increasing attention due to the popularity of 3D skeleton data. One main challenge lies in the large view variations in captured human actions. We propose a novel view adaptation scheme to automatically regulate observation viewpoints during the occurrence of an action. Rather than re-positioning the skeletons based on a human defined prior criterion, we design a view adaptive recurrent neural network (RNN) with LSTM architecture, which enables the network itself to adapt to the most suitable observation viewpoints from end to end. Extensive experiment analyses show that the proposed view adaptive RNN model strives to (1) transform the skeletons of various views to much more consistent viewpoints and (2) maintain the continuity of the action rather than transforming every frame to the same position with the same body orientation. Our model achieves significant improvement over the state-of-the-art approaches on three benchmark datasets.

****

Joint Discovery of Object States and Manipulation Actions

Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, Simon Lacoste-Julien; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2127-2136

Many human activities involve object manipulations aiming to modify the object state. Examples of common state changes include full/empty bottle, open/closed door, and attached/detached car wheel. In this work, we seek to automatically discover the states of objects and the associated manipulation actions. Given a set of videos for a particular task, we propose a joint model that learns to identify object states and to localize state-modifying actions. Our model is formulated as a discriminative clustering cost with constraints. We assume a consistent temporal order for the changes in object states and manipulation actions, and introduce new optimization techniques to learn model parameters without additional supervision. We demonstrate successful discovery of seven manipulation actions and corresponding object states on a new dataset of videos depicting real-life object manipulations. We show that our joint formulation results in an improvement of object state discovery by action recognition and vice versa.

****

What Actions Are Needed for Understanding Human Actions in Videos?

Gunnar A. Sigurdsson, Olga Russakovsky, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2137-2146

What is the right way to reason about human activities? What directions forward are most promising? In this work, we analyze the current state of human activity understanding in videos. The goal of this paper is to examine datasets, evaluation metrics, algorithms, and potential future directions. We look at the qualitative attributes that define activities such as pose variability, brevity, and density. The experiments consider multiple state-of-the-art algorithms and multiple datasets. The results demonstrate that while there is inherent ambiguity in the temporal extent of activities, current datasets still permit effective benchmarking. We discover that fine-grained understanding of objects and pose when combined with temporal reasoning is likely to yield substantial improvements in algorithmic accuracy. We present the many kinds of information that will be needed to achieve substantial gains in activity understanding: objects, verbs, intent, and sequential reasoning. The software and additional information will be made available to provide other researchers detailed diagnostics to understand their own algorithms.

****

Lattice Long Short-Term Memory for Human Action Recognition

Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E. Shi, Silvio Savarese; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2147-2156

Human actions captured in video sequences are three-dimensional signals characterizing visual appearance and motion dynamics. To learn action patterns, existing methods adopt Convolutional and/or Recurrent Neural Networks (CNNs and RNNs). CNN based methods are effective in learning spatial appearances, but are limited in modeling long-term motion dynamics. RNNs, especially Long Short-Term Memory (LSTM), are able to learn temporal motion dynamics. However, naively applying RNNs to video sequences in a convolutional manner implicitly assumes that motions in videos are stationary across different spatial locations. This assumption is valid for short-term motions but invalid when the duration of the motion is long.

In this work, we propose Lattice-LSTM, which extends LSTM by learning independent hidden state transitions of memory cells for individual spatial locations. This method effectively enhances the ability to model dynamics across time and addresses the non-stationary issue of long-term motion dynamics without significantly increasing the model complexity. Additionally, we introduce a novel multi-modal training procedure for training our network. Unlike traditional two-stream architectures which use RGB and optical flow information as input, our two-stream model leverages both modalities to jointly train both input gates and both forget gates in the network rather than treating the two streams as separate entities with no information about the other. We apply this end-to-end system to benchmark datasets (UCF-101 and HMDB-51) of human action recognition. Experiments show that on both datasets, our proposed method outperforms all existing ones that are based on LSTM and/or CNNs of similar model complexities.
****
Common Action Discovery and Localization in Unconstrained Videos
Jiong Yang, Junsong Yuan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2157-2166
Similar to common object discovery in images or videos, it is of great interests to discover and locate common actions in videos, which can benefit many video analytics applications such as video summarization, search, and understanding. In this work, we tackle the problem of common action discovery and localization in unconstrained videos, where we do not assume to know the types, numbers or locations of the common actions in the videos. Furthermore, each video can contain zero, one or several common action instances. To perform automatic discovery and localization in such challenging scenarios, we first generate action proposals using human prior. By building an affinity graph among all action proposals, we formulate the common action discovery as a subgraph density maximization problem to select the proposals containing common actions. To avoid enumerating in the exponentially large solution space, we propose an efficient polynomial time optimization algorithm. It solves the problem up to a user specified error bound with respect to the global optimal solution. The experimental results on several datasets show that even without any prior knowledge of common actions, our method can robustly locate the common actions in a collection of videos.
****
Pixel-Level Matching for Video Object Segmentation Using Convolutional Neural Networks
Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2167-2176
We propose a novel video object segmentation algorithm based on pixel-level matching using Convolutional Neural Networks (CNN). Our network aims to distinguish the target area from the background on the basis of the pixel-level similarity between two object units. The proposed network represents a target object using features from different depth layers in order to take advantage of both the spatial details and the category-level semantic information. Furthermore, we propose a feature compression technique that drastically reduces the memory requirements while maintaining the capability of feature representation. Two-stage training (pre-training and fine-tuning) allows our network to handle any target object regardless of its category (even if the object's type does not belong to the pre-training data) or of variations in its appearance through a video sequence. Experiments on large datasets demonstrate the effectiveness of our model - against related methods - in terms of accuracy, speed, and stability. Finally, we introduce the transferability of our network to different domains, such as the infrared data domain.
****
Am I a Baller? Basketball Performance Assessment From First-Person Videos
Gedas Bertasius, Hyun Soo Park, Stella X. Yu, Jianbo Shi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2177-2185
This paper presents a method to assess a basketball player's performance from his/her first-person video. A key challenge lies in the fact that the evaluation m

etric is highly subjective and specific to a particular evaluator. We leverage the first-person camera to address this challenge. The spatiotemporal visual semantics provided by a first-person view allows us to reason about the camera wearer's actions while he/she is participating in an unscripted basketball game. Our method takes a player's first-person video and provides a player's performance measure that is specific to an evaluator's preference. To achieve this goal, we first use a convolutional LSTM network to detect atomic basketball events from first-person videos. Our network's ability to zoom-in to the salient regions addresses the issue of a severe camera wearer's head movement in first-person videos. The detected atomic events are then passed through the Gaussian mixtures to construct a highly non-linear visual spatiotemporal basketball assessment feature. Finally, we use this feature to learn a basketball assessment model from pairs of labeled first-person basketball videos, for which a basketball expert indicates, which of the two players is better. We demonstrate that despite not knowing the basketball evaluator's criterion, our model learns to accurately assess the players in real-world games. Furthermore, our model can also discover basketball events that contribute positively and negatively to a player's performance.

****

Deep Cropping via Attention Box Prediction and Aesthetics Assessment
Wenguan Wang, Jianbing Shen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2186-2194

We model the photo cropping problem as a cascade of attention box regression and aesthetic quality classification, based on deep learning. A neural network is designed that has two branches for predicting attention bounding box and analyzing aesthetics, respectively. The predicted attention box is treated as an initial crop window where a set of cropping candidates are generated around it, without missing important information. Then, aesthetics assessment is employed to select the final crop as the one with the best aesthetic quality. With our network, cropping candidates share features within full-image convolutional feature maps, thus avoiding repeated feature computation and leading to higher computation efficiency. Via leveraging rich data for attention prediction and aesthetics assessment, the proposed method produces high-quality cropping results, even with the limited availability of training data for photo cropping. The experimental results demonstrate the competitive results and fast processing speed (5 fps with all steps).

****

Raster-To-Vector: Revisiting Floorplan Transformation
Chen Liu, Jiajun Wu, Pushmeet Kohli, Yasutaka Furukawa; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2195-2203

This paper addresses the problem of converting a rasterized floorplan image into a vector-graphics representation. Unlike existing approaches that rely on a sequence of low-level image processing heuristics, we adopt a learning-based approach. A neural architecture first transforms a rasterized image to a set of junctions that represent low-level geometric and semantic information (e.g., wall corners or door end-points). Integer programming is then formulated to aggregate junctions into a set of simple primitives (e.g., wall lines, door lines, or icon boxes) to produce a vectorized floorplan, while ensuring a topologically and geometrically consistent result. Our algorithm significantly outperforms existing methods and achieves around 90% precision and recall, getting to the range of production-ready performance. The vector representation allows 3D model popup for better indoor scene visualization, direct model manipulation for architectural remodeling, and further computational applications such as data analysis. Our system is efficient: we have converted hundred thousand production-level floorplan images into the vector representation and generated 3D popup models.

****

Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework
Michal Busta, Lukas Neumann, Jiri Matas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2204-2212

A method for scene text localization and recognition is proposed. The novelties include: training of both text detection and recognition in a single end-to-end pass, the structure of the recognition CNN and the geometry of its input layer that preserves the aspect of the text and adapts its resolution to the data. The proposed method achieves state-of-the-art accuracy in the end-to-end text recognition on two standard datasets - ICDAR 2013 and ICDAR 2015, whilst being an order of magnitude faster than competing methods - the whole pipeline runs at 10 frames per second on an NVidia K80 GPU.

****

Playing for Benchmarks

Stephan R. Richter, Zeeshan Hayder, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2213-2222

We present a benchmark suite for visual perception. The benchmark is based on more than 250K high-resolution video frames, all annotated with ground-truth data for both low-level and high-level vision tasks, including optical flow, semantic instance segmentation, object detection and tracking, object-level 3D scene layout, and visual odometry. Ground-truth data for all tasks is available for every frame. The data was collected while driving, riding, and walking a total of 184 kilometers in diverse ambient conditions in a realistic virtual world. To create the benchmark, we have developed a new approach to collecting ground-truth data from simulated worlds without access to their source code or content. We conduct statistical analyses that show that the composition of the scenes in the benchmark closely matches the composition of corresponding physical environments. The realism of the collected data is further validated via perceptual experiments. We analyze the performance of state-of-the-art methods for multiple tasks, providing reference baselines and highlighting challenges for future research.

****

Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223-2232

Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. We present an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. Our goal is to learn a mapping G: X -> Y such that the distribution of images from G(X) is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping F: Y -> X and introduce a   cycle consistency loss to push F(G(X)) ~ X (and vice versa). Qualitative results are presented on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach.

****

GANs for Biological Image Synthesis

Anton Osokin, Anatole Chessel, Rafael E. Carazo Salas, Federico Vaggi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2233-2242

In this paper, we propose a novel application of Generative Adversarial Networks (GAN) to the synthesis of cells imaged by fluorescence microscopy. Compared to natural images, cells tend to have a simpler and more geometric global structure that facilitates image generation. However, the correlation between the spatial pattern of different fluorescent proteins reflects important biological functions, and synthesized images have to capture these relationships to be relevant for biological applications. We adapt GANs to the task at hand and propose new models with casual dependencies between image channels that can generate multi-channel images, which would be impossible to obtain experimentally. We evaluate our approach using two independent techniques and compare it against sensible baselines. Finally, we demonstrate that by interpolating across the latent space we can mimic the known changes in protein localization that occur through time during the cell cycle, allowing us to predict temporal evolution from static images.

****

Learning to Synthesize a 4D RGBD Light Field From a Single Image

Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, Ren Ng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2243-2251

We present a machine learning algorithm that takes as input a 2D RGB image and synthesizes a 4D RGBD light field (color and depth of the scene in each ray direction). For training, we introduce the largest public light field dataset, consisting of over 3300 plenoptic camera light fields of scenes containing flowers and plants. Our synthesis pipeline consists of a convolutional neural network (CNN) that estimates scene geometry, a stage that renders a Lambertian light field using that geometry, and a second CNN that predicts occluded rays and non-Lambertian effects. Our algorithm builds on recent view synthesis methods, but is unique in predicting RGBD for each light field ray and improving unsupervised single image depth estimation by enforcing consistency of ray depths that should intersect the same scene point.

****

Neural EPI-Volume Networks for Shape From Light Field

Stefan Heber, Wei Yu, Thomas Pock; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2252-2260

This paper presents a novel deep regression network to extract geometric information from Light Field (LF) data. Our network builds upon u-shaped network architectures. Those networks involve two symmetric parts, an encoding and a decoding part. In the first part the network encodes relevant information from the given input into a set of high-level feature maps. In the second part the generated feature maps are then decoded to the desired output. To predict reliable and robust depth information the proposed network examines 3D subsets of the 4D LF called Epipolar Plane Image (EPI) volumes. An important aspect of our network is the use of 3D convolutional layers, that allow to propagate information from two spatial dimensions and one directional dimension of the LF. Compared to previous work this allows for an additional spatial regularization, which reduces depth artifacts and simultaneously maintains clear depth discontinuities. Experimental results show that our approach allows to create high-quality reconstruction results, which outperform current state-of-the-art Shape from Light Field (SfLF) techniques. The main advantage of the proposed approach is the ability to provide those high-quality reconstructions at a low computation time.

****

Material Editing Using a Physically Based Rendering Network

Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, Jyh-Ming Lien; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2261-2269

The ability to edit materials of objects in images is desirable by many content creators. However, this is an extremely challenging task as it requires to disentangle intrinsic physical properties of an image. We propose an end-to-end network architecture that replicates the forward image formation process to accomplish this task. Specifically, given a single image, the network first predicts intrinsic properties, i.e. shape, illumination, and material, which are then provided to a rendering layer. This layer performs in-network image synthesis, thereby enabling the network to understand the physics behind the image formation process. The proposed rendering layer is fully differentiable, supports both diffuse and specular materials, and thus can be applicable in a variety of problem settings. We demonstrate a rich set of visually plausible material editing examples and provide an extensive comparative study.

****

Turning Corners Into Cameras: Principles and Methods

Katherine L. Bouman, Vickie Ye, Adam B. Yedidia, Fredo Durand, Gregory W. Wornell, Antonio Torralba, William T. Freeman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2270-2278

We show that walls and other obstructions with edges can be exploited as naturally-occurring "cameras" that reveal the hidden scenes beyond them. In particular, we demonstrate methods for using the subtle spatio-temporal radiance variations that arise on the ground at the base of edges to construct a one-dimensional video of the hidden scene. The resulting technique can be used for a variety of applications in diverse physical settings. From standard RGB video recordings of the variations in intensity, we use edge cameras to recover a 1-D video that reveals the number and trajectories of people moving in an occluded scene. We further show that adjacent vertical edges, such as those that arise in the case of an open doorway, yield a stereo camera from which the 2-D location of hidden, moving objects can be recovered. We demonstrate our technique in a number of indoor and outdoor environments involving varied surfaces and illumination conditions.

****

Linear Differential Constraints for Photo-Polarimetric Height Estimation

Silvia Tozza, William A. P. Smith, Dizhong Zhu, Ravi Ramamoorthi, Edwin R. Hancock; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2279-2287

In this paper we present a differential approach to photo-polarimetric shape estimation. We propose several alternative differential constraints based on polarisation and photometric shading information and show how to express them in a unified partial differential system. Our method uses the image ratios technique to combine shading and polarisation information in order to directly reconstruct surface height, without first computing surface normal vectors. Moreover, we are able to remove the non-linearities so that the problem reduces to solving a linear differential problem. We also introduce a new method for estimating a polarisation image from multichannel data and, finally, we show it is possible to estimate the illumination directions in a two source setup, extending the method into an uncalibrated scenario. From a numerical point of view, we use a least-squares formulation of the discrete version of the problem. To the best of our knowledge, this is the first work to consider a unified differential approach to solve photo-polarimetric shape estimation directly for height. Numerical results on synthetic and real-world data confirm the effectiveness of our proposed method.

****

Polynomial Solvers for Saturated Ideals

Viktor Larsson, Kalle Astrom, Magnus Oskarsson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2288-2297

In this paper we present a new method for creating polynomial solvers for problems where a (possibly infinite) subset of the solutions are undesirable or uninteresting. These solutions typically arise from simplifications made during modeling, but can also come from degeneracies which are inherent to the geometry of the original problem. The proposed approach extends the standard action matrix method to saturated ideals. This allows us to add constraints that some polynomials should be non-zero on the solutions. This does not only offer the possibility of improved performance by removing superfluous solutions, but makes a larger class of problems tractable. Previously, problems with infinitely many solutions could not be solved directly using the action matrix method as it requires a zero-dimensional ideal. In contrast we only require that after removing the unwanted solutions only finitely many remain. We evaluate our method on three applications, optimal triangulation, time-of-arrival self-calibration and optimal vanishing point estimation.

****

Shape Inpainting Using 3D Generative Adversarial Network and Recurrent Convolutional Networks

Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, Ulrich Neumann; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2298-2306

Recent advances in convolutional neural networks have shown promising results in 3D shape completion. But due to GPU memory limitations, these methods can only produce low-resolution outputs. To inpaint 3D models with semantic plausibility and contextual details, we introduce a hybrid framework that combines a 3D Encoder-Decoder Generative Adversarial Network (3D-ED-GAN) and a Long-term Recurrent Convolutional Network (LRCN). The 3D-ED-GAN is a 3D convolutional neural network trained with a generative adversarial paradigm to fill missing 3D data in low-r esolution. LRCN adopts a recurrent neural network architecture to minimize GPU memory usage and incorporates an Encoder-Decoder pair into a Long Short-term Memory Network. By handling the 3D model as a sequence of 2D slices, LRCN transforms a coarse 3D shape into a more complete and higher resolution volume. While 3D-ED-GAN captures global contextual structure of the 3D shape, LRCN localizes the fine-grained details. Experimental results on both real-world and synthetic data show reconstructions from corrupted models result in complete and high-resolution 3D objects.

****

SurfaceNet: An End-To-End 3D Neural Network for Multiview Stereopsis

Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, Lu Fang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2307-2315

This paper proposes an end-to-end learning framework for multiview stereopsis. We term the network SurfaceNet. It takes a set of images and their corresponding camera parameters as input and directly infers the 3D model. The key advantage of the framework is that both photo-consistency as well geometric relations of the surface structure can be directly learned for the purpose of multiview stereopsis in an end-to-end fashion. SurfaceNet is a fully 3D convolutional network which is achieved by encoding the camera parameters together with the images in a 3D voxel representation. We evaluate SurfaceNet on the large-scale DTU benchmark.

****

Making Minimal Solvers for Absolute Pose Estimation Compact and Robust

Viktor Larsson, Zuzana Kukelova, Yinqiang Zheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2316-2324

In this paper we present new techniques for constructing compact and robust minimal solvers for absolute pose estimation. We focus on the P4Pfr problem, but the methods we propose are applicable to a more general setting. Previous approaches to P4Pfr suffer from artificial degeneracies which come from their formulation and not the geometry of the original problem. In this paper we show how to avoid these false degeneracies to create more robust solvers. Combined with recently published techniques for Grobner basis solvers we are also able to construct solvers which are significantly smaller. We evaluate our solvers on both real and synthetic data, and show improved performance compared to competing solvers. Finally we show that our techniques can be directly applied to the P3.5Pf problem to get a non-degenerate solver, which is competitive with the current state-of-the-art.

****

3D Surface Detail Enhancement From a Single Normal Map

Wuyuan Xie, Miaohui Wang, Xianbiao Qi, Lei Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2325-2333

In 3D reconstruction, the obtained surface details are mainly limited to the visual sensor due to sampling and quantization in the digitalization process. How to get a fine-grained 3D surface with low-cost is still a challenging obstacle in terms of experience, equipment and easy-to-obtain. This work introduces a novel framework for enhancing surfaces reconstructed from normal map, where the assumptions on hardware (e.g., photometric stereo setup) and reflection model (e.g., Lambertion reflection) are not necessarily needed. We propose to use a new measure, angle profile, to infer the hidden micro-structure from existing surfaces. In addition, the inferred results are further improved in the domain of discrete geometry processing (DGP) which is able to achieve a stable surface structure under a selectable enhancement setting. Extensive simulation results show that the proposed method obtains significantly improvements over uniform sharpening method in terms of both subjective visual assessment and objective quality metric.

****

RMPE: Regional Multi-Person Pose Estimation

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2334-2343

Multi-person pose estimation in the wild is challenging. Although state-of-the-art human detectors have demonstrated good performance, small errors in localization and recognition are inevitable. These errors can cause failures for a single

-person pose estimator (SPPE), especially for methods that solely depend on human detection results. In this paper, we propose a novel regional multi-person pose estimation (RMPE) framework to facilitate pose estimation in the presence of inaccurate human bounding boxes. Our framework consists of three components: Symmetric Spatial Transformer Network (SSTN), Parametric Pose Non-Maximum-Suppression (NMS), and Pose-Guided Proposals Generator (PGPG). Our method is able to handle inaccurate bounding boxes and redundant detections, allowing it to achieve 76.7 mAP on the MPII (multi person) dataset. Our model and source codes are made publicly available.

****

Online Video Object Detection Using Association LSTM

Yongyi Lu, Cewu Lu, Chi-Keung Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2344-2352

Video object detection is a fundamental tool for many applications. Since direct application of image-based object detection cannot leverage the rich temporal information inherent in video data, we advocate to the detection of long-range video object pattern. While the Long Short-Term Memory (LSTM) has been the de facto choice for such detection, currently LSTM cannot fundamentally model object association between consecutive frames. In this paper, we propose the association LSTM to address this fundamental association problem. Association LSTM not only regresses and classifiy directly on object locations and categories but also associates features to represent each output object. By minimizing the matching error between these features, we learn how to associate objects in two consecutive frames. Additionally, our method works in an online manner, which is important for most video tasks. Compared to the traditional video object detection methods, our approach outperforms them on standard video datasets.

****

PolyFit: Polygonal Surface Reconstruction From Point Clouds

Liangliang Nan, Peter Wonka; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2353-2361

We propose a novel framework for reconstructing lightweight polygonal surfaces from point clouds. Unlike traditional methods that focus on either extracting good geometric primitives or obtaining proper arrangements of primitives, the emphasis of this work lies in intersecting the primitives (planes only) and seeking for an appropriate combination of them to obtain a manifold polygonal surface model without boundary. We show that reconstruction from point clouds can be cast as a binary labeling problem. Our method is based on a hypothesizing and selection strategy. We first generate a reasonably large set of face candidates by intersecting the extracted planar primitives. Then an optimal subset of the candidate faces is selected through optimization. Our optimization is based on a binary linear programming formulation under hard constraints that enforce the final polygonal surface model to be manifold and watertight. Experiments on point clouds from various sources demonstrate that our method can generate lightweight polygonal surface models of arbitrary piecewise planar objects. Besides, our method is capable of recovering sharp features and is robust to noise, outliers, and missing data.

****

Progressive Large Scale-Invariant Image Matching in Scale Space

Lei Zhou, Siyu Zhu, Tianwei Shen, Jinglu Wang, Tian Fang, Long Quan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2362-2371

The power of modern image matching approaches is still fundamentally limited by the abrupt scale changes in images. In this paper, we propose a scale-invariant image matching approach to tackling the very large scale variation of views. Drawing inspiration from the scale space theory, we start with encoding the image's scale space into a compact multi-scale representation. Then, rather than trying to find the exact feature matches all in one step, we propose a progressive two-stage approach. First, we determine the related scale levels in scale space, enclosing the inlier feature correspondences, based on an optimal and exhaustive matching in a limited scale space. Second, we produce both the image similarity m easurement and feature correspondences simultaneously after restricting matching between the related scale levels in a robust way. The matching performance has been intensively evaluated on vision tasks including image retrieval, feature matching and Structure-from-Motion (SfM). The successful integration of the challenging fusion of high aerial and low ground-level views with significant scale differences manifests the superiority of the proposed approach.

****

Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map

Liu Liu, Hongdong Li, Yuchao Dai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2372-2381

Given an image of a street scene in a city, this paper develops a new method that can quickly and precisely pinpoint at which location (as well as viewing direction) the image was taken, against a pre-stored large-scale 3D point-cloud map of the city. We adopt the recently developed 2D-3D direct feature matching framework for this task [23,31,32,42-44]. This is a challenging task especially for large-scale problems. As the map size grows bigger, many 3D points in the wider geographical area can be visually very similar-or even identical-causing severe ambiguities in 2D-3D feature matching. The key is to quickly and unambiguously find the correct matches between a query image and the large 3D map. Existing methods solve this problem mainly via comparing individual features' visual similarities in a local and per feature manner, thus only local solutions can be found, inadequate for large-scale applications. In this paper, we introduce a global method which harnesses global contextual information exhibited both within the query image and among all the 3D points in the map. This is achieved by a novel global ranking algorithm, applied to a Markov network built upon the 3D map, which takes account of not only visual similarities between individual 2D-3D matches, but also their global compatibilities (as measured by co-visibility) among all matching pairs found in the scene. Tests on standard benchmark datasets show that our method achieved both higher precision and comparable recall, compared with the state-of-the-art.

****

Multi-View Non-Rigid Refinement and Normal Selection for High Quality 3D Reconstruction

Sk. Mohammadul Haque, Venu Madhav Govindu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2382-2390

In recent years, there have been a variety of proposals for high quality 3D reconstruction by fusion of depth and normal maps that contain good low and high frequency information respectively. Typically, these methods create an initial mesh representation of the complete object or scene being scanned. Subsequently, normal estimates are assigned to each mesh vertex and a mesh-normal fusion step is carried out. In this paper, we present a complete pipeline for such depth-normal fusion. The key innovations in our pipeline are twofold. Firstly, we introduce a global multi-view non-rigid refinement step that corrects for the non-rigid misalignment present in the depth and normal maps. We demonstrate that such a correction is crucial for preserving fine-scale 3D features in the final reconstruction. Secondly, despite adequate care, the averaging of multiple normals invariably results in blurring of 3D detail. To mitigate this problem, we propose an approach that selects one out of many available normals. Our global cost for normal selection incorporates a variety of desirable properties and can be efficiently solved using graph cuts. We demonstrate the efficacy of our approach in generating high quality 3D reconstructions of both synthetic and real 3D models and compare with existing methods in the literature.

****

Multi-Stage Multi-Recursive-Input Fully Convolutional Networks for Neuronal Boundary Detection

Wei Shen, Bin Wang, Yuan Jiang, Yan Wang, Alan Yuille; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2391-2400

In the field of connectomics, neuroscientists seek to identify cortical connectivity comprehensively. Neuronal boundary detection from the Electron Microscopy (EM) images is often done to assist the automatic reconstruction of neuronal circ uit. But the segmentation of EM images is a challenging problem, as it requires the detector to be able to detect both filament-like thin and blob-like thick membrane, while suppressing the ambiguous intracellular structure. In this paper, we propose multi-stage multi-recursiveinput fully convolutional networks to address this problem. The multiple recursive inputs for one stage, i.e., the multiple side outputs with different receptive field sizes learned from the lower stage, provide multi-scale contextual boundary information for the consecutive learning. This design is biologically-plausible, as it likes a human visual system to compare different possible segmentation solutions to address the ambiguous boundary issue. Our multi-stage networks are trained end-to-end. It achieves promising results on two public available EM segmentation datasets, the mouse piriform cortex dataset and the ISBI 2012 EM dataset.

****

Depth and Image Restoration From Light Field in a Scattering Medium

Jiandong Tian, Zachary Murez, Tong Cui, Zhen Zhang, David Kriegman, Ravi Ramamoorthi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2401-2410

Traditional imaging methods and computer vision algorithms are often ineffective when images are acquired in scattering media, such as underwater, fog, and biological tissue. Here, we explore the use of light field imaging and algorithms for image restoration and depth estimation that address the image degradation from the medium. Towards this end, we make the following three contributions. First, we present a new single image restoration algorithm which removes backscatter and attenuation from images better than existing methods, and apply it to each view in the light field. Second, we combine a novel transmission based depth cue with existing correspondence and defocus cues to improve light field depth estimation. In densely scattering media, our transmission depth cue is critical for depth estimation since the images have low signal to noise ratios which significantly degrades the performance of the correspondence and defocus cues. Finally, we propose shearing and refocusing multiple views of the light field to recover a single image of higher quality than what is possible from a single view. We demonstrate the benefits of our method through extensive experimental results in a water tank.

****

Video Reflection Removal Through Spatio-Temporal Optimization

Ajay Nandoriya, Mohamed Elgharib, Changil Kim, Mohamed Hefeeda, Wojciech Matusik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2411-2419

Reflections can obstruct content during video capture and hence their removal is desirable. Current removal techniques are designed for still images, extracting only one reflection (foreground) and one background layer from the input. When extended to videos, unpleasant artifacts such as temporal flickering and incomplete separation are generated. We present a technique for video reflection removal by jointly solving for motion and separation. The novelty of our work is in our optimization formulation as well as the motion initialization strategy. We present a novel spatio-temporal optimization that takes n frames as input and directly estimates 2n frames as output, n for each layer. We aim to fully utilize spatio-temporal information in our objective terms. Our motion initialization is based on iterative frame-to-frame alignment instead of the direct alignment used by current approaches. We compare against advanced video extensions of the state of the art, and we significantly reduce temporal flickering and improve separation. In addition, we reduce image blur and recover moving objects more accurately. We validate our approach through subjective and objective evaluations on real and controlled data.

****

Efficient Online Local Metric Adaptation via Negative Samples for Person Re-Identification

Jiahuan Zhou, Pei Yu, Wei Tang, Ying Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2420-2428

Many existing person re-identification (PRID) methods typically attempt to train a faithful global metric offline to cover the enormous visual appearance variations, so as to directly use it online on various probes for identity matching. However, their need for a huge set of positive training pairs is very demanding in practice. In contrast to these methods, this paper advocates a different paradigm: part of the learning can be performed online but with nominal costs, so as to achieve online metric adaptation for different input probes. A major challenge here is that no positive training pairs are available for the probe anymore. By only exploiting easily-available negative samples, we propose a novel solution to achieve local metric adaptation effectively and efficiently. For each probe at the test time, it learns a strictly positive semi-definite dedicated local metric. Comparing to offline global metric learning, its computational cost is negligible. The insight of this new method is that the local hard negative samples can actually provide tight constraints to fine tune the metric locally. This new local metric adaptation method is generally applicable, as it can be used on top of any global metric to enhance its performance. In addition, this paper gives in-depth theoretical analysis and justification of the new method. We prove that our new method guarantees the reduction of the classification error asymptotically, and prove that it actually learns the optimal local metric to best approximate the asymptotic case by a finite number of training data. Extensive experiments and comparative studies on almost all major benchmarks (VIPeR, QMUL GRID, CUHK Campus, CUHK03 and Market-1501) have confirmed the effectiveness and superiority of our method.

****

Stepwise Metric Promotion for Unsupervised Video Person Re-Identification

Zimo Liu, Dong Wang, Huchuan Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2429-2438

The intensive annotation cost and the rich but unlabeled data contained in videos motivate us to propose an unsupervised video-based person re-identification (re-ID) method. We start from two assumptions: 1) different video tracklets typically contain different persons, given that the tracklets are taken at distinct places or with long intervals; 2) within each tracklet, the frames are mostly of the same person. Based on these assumptions, this paper propose a stepwise metric promotion approach to estimate the identities of training tracklets, which iterates between cross-camera tracklet association and feature learning. Specifically, We use each training tracklet as a query, and perform retrieval in the cross camera training set. Our method is built on reciprocal nearest neighbor search and can eliminate the hard negative label matches, i.e., the cross-camera nearest neighbors of the false matches in the initial rank list. The tracklet that passes the reciprocal nearest neighbor check is considered to have the same ID with the query. Experimental results on the PRID 2011, ILIDS-VID, and MARS datasets show that the proposed method achieves very competitive re-ID accuracy compared with its supervised counterparts.

****

Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis

Rui Huang, Shu Zhang, Tianyu Li, Ran He; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2439-2448

Photorealistic frontal view synthesis from a single face image has a wide range of applications in the field of face recognition. Although data-driven deep learning methods have been proposed to address this problem by seeking solutions from ample face data, this problem is still challenging because it is intrinsically ill-posed. This paper proposes a Two-Pathway Generative Adversarial Network (TP-GAN) for photorealistic frontal view synthesis by simultaneously perceiving global structures and local details. Four landmark located patch networks are proposed to attend to local textures in addition to the commonly used global encoder-decoder network. Except for the novel architecture, we make this ill-posed problem well constrained by introducing a combination of adversarial loss, symmetry loss and identity preserving loss. The combined loss function leverages both frontal face distribution and pre-trained discriminative deep face models to guide an identity preserving inference of frontal views from profiles. Different from p revious deep learning methods that mainly rely on intermediate features for recognition, our method directly leverages the synthesized identity preserving image for downstream tasks like face recognition and attribution estimation. Experimental results demonstrate that our method not only presents compelling perceptual results but also outperforms state-of-the-art results on large pose face recognition.

****

Group Re-Identification via Unsupervised Transfer of Sparse Features Encoding

Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, Gian Luca Foresti; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2449-2458

Person re-identification is best known as the problem of associating a single person that is observed from one or more disjoint cameras. The existing literature has mainly addressed such an issue, neglecting the fact that people usually move in groups, like in crowded scenarios. We believe that the additional information carried by neighboring individuals provides a relevant visual context that can be exploited to obtain a more robust match of single persons within the group. Despite this, re-identifying groups of people compound the common single person re-identification problems by introducing changes in the relative position of persons within the group and severe self-occlusions. In this paper, we propose a solution for group re-identification that grounds on transferring knowledge from single person re-identification to group re-identification by exploiting sparse dictionary learning. First, a dictionary of sparse atoms is learned using patches extracted from single person images. Then, the learned dictionary is exploited to obtain a sparsity-driven residual group representation, which is finally matched to perform the re-identification. Extensive experiments on the i-LIDS groups and two newly collected datasets show that the proposed solution outperforms state-of-the-art approaches.

****

Visual Transformation Aided Contrastive Learning for Video-Based Kinship Verification

Hamdi Dibeklioglu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2459-2468

Automatic kinship verification from facial information is a relatively new and open research problem in computer vision. This paper explores the possibility of learning an efficient facial representation for video-based kinship verification by exploiting the visual transformation between facial appearance of kin pairs. To this end, a Siamese-like coupled convolutional encoder-decoder network is proposed. To reveal resemblance patterns of kinship while discarding the similarity patterns that can also be observed between people who do not have a kin relationship, a novel contrastive loss function is defined in the visual appearance space. For further optimization, the learned representation is fine-tuned using a feature-based contrastive loss. An expression matching procedure is employed in the model to minimize the negative influence of expression differences between kin pairs. Each kin video is analyzed by a sliding temporal window to leverage short-term facial dynamics. The effectiveness of the proposed method is assessed on seven different kin relationships using smile videos of kin pairs. On the average, 93.65% verification accuracy is achieved, improving the state of the art.

****

Decoder Network Over Lightweight Reconstructed Feature for Fast Semantic Style Transfer

Ming Lu, Hao Zhao, Anbang Yao, Feng Xu, Yurong Chen, Li Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2469-2477

Recently, the community of style transfer is trying to incorporate semantic information into traditional system. This practice achieves better perceptual results by transferring the style between semantically-corresponding regions. Yet, few efforts are invested to address the computation bottleneck of back-propagation. In this paper, we propose a new framework for fast semantic style transfer. Our method decomposes the semantic style transfer problem into feature reconstruction part and feature decoder part. The reconstruction part tactfully solves the o optimization problem of content loss and style loss in feature space by particularly reconstructed feature. This significantly reduces the computation of propagating the loss through the whole network. The decoder part transforms the reconstructed feature into the stylized image. Through a careful bridging of the two modules, the proposed approach not only achieves competitive results as backward optimization methods but also is about two orders of magnitude faster.

****

Blind Image Deblurring With Outlier Handling

Jiangxin Dong, Jinshan Pan, Zhixun Su, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2478-2486

Deblurring images with outliers has attracted considerable attention recently. However, existing algorithms usually involve complex operations which increase the difficulty of blur kernel estimation. In this paper, we propose a simple yet effective blind image deblurring algorithm to handle blurred images with outliers. The proposed method is motivated by the observation that outliers in the blurred images significantly affect the goodness-of-fit in function approximation. Therefore, we propose an algorithm to model the data fidelity term so that the outliers have little effect on kernel estimation. The proposed algorithm does not require any heuristic outlier detection step, which is critical to the state-of-the-art blind deblurring methods for images with outliers. We analyze the relationship between the proposed algorithm and other blind deblurring methods with outlier handling and show how to estimate intermediate latent images for blur kernel estimation principally. We show that the proposed method can be applied to generic image deblurring as well as non-uniform deblurring. Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art blind image deblurring methods on both synthetic and real-world images.

****

Paying Attention to Descriptions Generated by Image Captioning Models

Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, Jorma Laaksonen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2487-2496

To bridge the gap between humans and machines in image understanding and describing, we need further insight into how people describe a perceived scene. In this paper, we study the agreement between bottom-up saliency-based visual attention and object referrals in scene description constructs. We investigate the properties of human-written descriptions and machine-generated ones. We then propose a saliency-boosted image captioning model in order to investigate benefits from low-level cues in language models. We learn that (1) humans mention more salient objects earlier than less salient ones in their descriptions, (2) the better a captioning model performs, the better attention agreement it has with human descriptions, (3) the proposed saliency-boosted model, compared to its baseline form, does not improve significantly on the MS COCO database, indicating explicit bottom-up boosting does not help when the task is well learnt and tuned on a data, (4) a better generalization is, however, observed for the saliency-boosted model on unseen data.

****

Fast Image Processing With Fully-Convolutional Networks

Qifeng Chen, Jia Xu, Vladlen Koltun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2497-2506

We present an approach to accelerating a wide variety of image processing operators. Our approach uses a fully-convolutional network that is trained on input-output pairs that demonstrate the operator's action. After training, the original operator need not be run at all. The trained network operates at full resolution and runs in constant time. We investigate the effect of network architecture on approximation accuracy, runtime, and memory footprint, and identify a specific architecture that balances these considerations. We evaluate the presented approach on ten advanced image processing operators, including multiple variational models, multiscale tone and detail manipulation, photographic style transfer, nonlocal dehazing, and nonphotorealistic stylization. All operators are approximated by the same model. Experiments demonstrate that the presented approach is significantly more accurate than prior approximation schemes. It increases approxima tion accuracy as measured by PSNR across the evaluated operators by 8.5 dB on the MIT-Adobe dataset (from 27.5 to 36 dB) and reduces DSSIM by a multiplicative factor of 3 compared to the most accurate prior approximation scheme, while being the fastest. We show that our models generalize across datasets and across resolutions, and investigate a number of extensions of the presented approach.

****

Robust Video Super-Resolution With Learned Temporal Dynamics

Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, Thomas Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2507-2515

Video super-resolution (SR) aims to generate a high-resolution (HR) frame from multiple low-resolution (LR) frames. The inter-frame temporal relation is as crucial as the intra-frame spatial relation for tackling this problem. However, how to utilize temporal information efficiently and effectively remains challenging since complex motion is difficult to model and can introduce adverse effects if not handled properly. We address this problem from two aspects. First, we propose a temporal adaptive neural network that can adaptively determine the optimal scale of temporal dependency. Filters on various temporal scales are applied to the input LR sequence before their responses are adaptively aggregated. Second, we reduce the complexity of motion between neighboring frames using a spatial alignment network that is much more robust and efficient than competing alignment methods and can be jointly trained with the temporal adaptive network in an end-to-end manner. Our proposed models with learned temporal dynamics are systematically evaluated on public video datasets and achieve state-of-the-art SR results compared with other recent video SR approaches. Both of the temporal adaptation and the spatial alignment modules are demonstrated to considerably improve SR quality over their plain counterparts.

****

Should We Encode Rain Streaks in Video as Deterministic or Stochastic?

Wei Wei, Lixuan Yi, Qi Xie, Qian Zhao, Deyu Meng, Zongben Xu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2516-2525

Videos taken in the wild sometimes contain unexpected rain streaks, which brings difficulty in subsequent video processing tasks. Rain streak removal in a video (RSRV) is thus an important issue and has been attracting much attention in computer vision. Different from previous RSRV methods formulating rain streaks as a deterministic message, this work first encodes the rains in a stochastic manner, i.e., a patch-based mixture of Gaussians. Such modification makes the proposed model capable of finely adapting a wider range of rain variations instead of certain types of rain configurations as traditional. By integrating with the spatiotemporal smoothness configuration of moving objects and low-rank structure of background scene, we propose a concise model for RSRV, containing one likelihood term imposed on the rain streak layer and two prior terms on the moving object and background scene layers of the video. Experiments implemented on videos with synthetic and real rains verify the superiority of the proposed method, as compared with the state-of-the-art methods, both visually and quantitatively in various performance metrics.

****

Joint Bi-Layer Optimization for Single-Image Rain Streak Removal

Lei Zhu, Chi-Wing Fu, Dani Lischinski, Pheng-Ann Heng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2526-2534

We present a novel method for removing rain streaks from a single input image by decomposing it into a rain-free background layer B and a rain-streak layer R. A joint optimization process is used that alternates between removing rain-streak details from B and removing non-streak details from R. The process is assisted by three novel image priors. Observing that rain streaks typically span a narrow range of directions, we first analyze the local gradient statistics in the rain image to identify image regions that are dominated by rain streaks. From these regions, we estimate the dominant rain streak direction and extract a collection of rain-dominated patches. Next, we define two priors on the background layer B, one based on a centralized sparse representation and another based on the esti mated rain direction. A third prior is defined on the rain-streak layer R, based on similarity of patches to the extracted rain patches. Both visual and quantitative comparisons demonstrate that our method outperforms the state-of-the-art.

****

Low-Dimensionality Calibration Through Local Anisotropic Scaling for Robust Hand Model Personalization

Edoardo Remelli, Anastasia Tkach, Andrea Tagliasacchi, Mark Pauly; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2535-2543

We present a robust algorithm for personalizing a sphere-mesh tracking model to a user from a collection of depth measurements. Our core contribution is to demonstrate how simple geometric reasoning can be exploited to build a shape-space, and how its performance is comparable to shape-spaces constructed from datasets of carefully calibrated models. We achieve this goal by first re-parameterizing the geometry of the tracking template, and introducing a multi-stage calibration optimization. Our novel parameterization decouples the degrees of freedom for pose and shape, resulting in improved convergence properties. Our analytically differentiable multi-stage calibration pipeline optimizes for the model in the natural low-dimensional space of local anisotropic scalings, leading to an effective solution that can be easily embedded in other tracking/calibration algorithms. Compared to existing sphere-mesh calibration algorithms, quantitative experiments assess our algorithm possesses a larger convergence basin, and our personalized models allows to perform motion tracking with superior accuracy. Code and data are available at http://github.com/edoRemelli/hadjust

****

Non-Markovian Globally Consistent Multi-Object Tracking

Andrii Maksai, Xinchao Wang, Francois Fleuret, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2544-2554

Many state-of-the-art approaches to multi-object tracking rely on detecting them in each frame independently, grouping detections into short but reliable trajectory segments, and then further grouping them into full trajectories. This grouping typically relies on imposing local smoothness constraints but almost never on enforcing more global ones on the trajectories. In this paper, we propose a non-Markovian approach to imposing global consistency by using behavioral patterns to guide the tracking algorithm. When used in conjunction with state-of-the-art tracking algorithms, this further increases their already good performance on multiple challenging datasets. We show significant improvements both in supervised settings where ground truth is available and behavioral patterns can be learned from it, and in completely unsupervised settings.

****

CREST: Convolutional Residual Learning for Visual Tracking

Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson W. H. Lau, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2555-2564

Discriminative correlation filters (DCFs) have \ryn been shown to perform superiorly in visual tracking. They \ryn only need a small set of training samples from the initial frame to generate an appearance model. However, existing DCFs learn the filters separately from feature extraction, and update these filters using a moving average operation with an empirical weight. These DCF trackers hardly benefit from the end-to-end training. In this paper, we propose the CREST algorithm to reformulate DCFs as a one-layer convolutional neural network. Our method integrates feature extraction, response map generation as well as model update into the neural networks for an end-to-end training. To reduce model degradation during online update, we apply residual learning to take appearance changes into account. Extensive experiments on the benchmark datasets demonstrate that our CREST tracker performs favorably against state-of-the-art trackers.

****

Volumetric Flow Estimation for Incompressible Fluids Using the Stationary Stokes Equations

Katrin Lasinger, Christoph Vogel, Konrad Schindler; Proceedings of the IEEE Inte rnational Conference on Computer Vision (ICCV), 2017, pp. 2565-2573

In experimental fluid dynamics, the flow in a volume of fluid is observed by injecting high-contrast tracer particles and tracking them in multi-view video. Fluid dynamics researchers have developed variants of space-carving to reconstruct the 3D particle distribution at a given time-step, and then use relatively simple local matching to recover the motion over time. On the contrary, estimating the optical flow between two consecutive images is a long-standing standard problem in computer vision, but only little work exists about volumetric 3D flow. Here, we propose a variational method for 3D fluid flow estimation from multi-view data. We start from a 3D version of the standard variational flow model, and investigate different regularization schemes that ensure divergence-free flow fields, to account for the physics of incompressible fluids. Moreover, we propose a semi-dense formulation, to cope with the computational demands of large volumetric datasets. Flow is estimated and regularized at a lower spatial resolution, while the data term is evaluated at full resolution to preserve the discriminative power and geometric precision of the local particle distribution. Extensive experiments reveal that a simple sum of squared differences (SSD) is the most suitable data term for our application. For regularization, an energy whose Euler-Lagrange equations correspond to the stationary Stokes equations leads to the best results. This strictly enforces a divergence-free flow and additionally penalizes the squared gradient of the flow.

****

Bounding Boxes, Segmentations and Object Coordinates: How Important Is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?

Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, Andreas Geiger; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2574-2583

Existing methods for 3D scene flow estimation often fail in the presence of large displacement or local ambiguities, e.g., at texture-less or reflective surfaces. However, these challenges are omnipresent in dynamic road scenes, which is the focus of this work. Our main contribution is to overcome these 3D motion estimation problems by exploiting recognition. In particular, we investigate the importance of recognition granularity, from coarse 2D bounding box estimates over 2D instance segmentations to fine-grained 3D object part predictions. We compute these cues using CNNs trained on a newly annotated dataset of stereo images and integrate them into a CRF-based model for robust 3D scene flow estimation - an approach we term Instance Scene Flow. We analyze the importance of each recognition cue in an ablation study and observe that the instance segmentation cue is by far strongest, in our setting. We demonstrate the effectiveness of our method on the challenging KITTI 2015 scene flow benchmark where we achieve state-of-the-art performance at the time of submission.

****

Performance Guaranteed Network Acceleration via High-Order Residual Quantization

Zefan Li, Bingbing Ni, Wenjun Zhang, Xiaokang Yang, Wen Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2584-2592

Input binarization has shown to be an effective way for network acceleration. However, previous binarization scheme could be regarded as simple pixel-wise thresholding operations (i.e., order-one approximation) and suffers a big accuracy loss. In this paper, we propose a high-order binarization scheme, which achieves more accurate approximation while still possesses the advantage of binary operation. In particular, the proposed scheme recursively performs residual quantization and yields a series of binary input images with decreasing magnitude scales. Accordingly, we propose high-order binary filtering and gradient propagation operations for both forward and backward computations. Theoretical analysis shows approximation error guarantee property of proposed method. Extensive experimental results demonstrate that the proposed scheme yields great recognition accuracy while being accelerated.

****

Deep Metric Learning With Angular Loss

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, Yuanqing Lin; Proceedings of the IEE

E International Conference on Computer Vision (ICCV), 2017, pp. 2593-2601

The modern image search system requires semantic understanding of image, and a key yet under-addressed problem is to learn a good metric for measuring the similarity between images. While deep metric learning has yielded impressive performance gains by extracting high level abstractions from image data, a proper objective loss function becomes the central issue to boost the performance. In this paper, we propose a novel angular loss, which takes angle relationship into account, for learning better similarity metric. Whereas previous metric learning methods focus on optimizing the similarity (contrastive loss) or relative similarity (triplet loss) of image pairs, our proposed method aims at constraining the angle at the negative point of triplet triangles. Several favorable properties are observed when compared with conventional methods. First, scale invariance is introduced, improving the robustness of objective against feature variance. Second, a third-order geometric constraint is inherently imposed, capturing additional local structure of triplet triangles than contrastive loss or triplet loss. Third, better convergence has been demonstrated by experiments on three publicly available datasets.

****

Compositional Human Pose Regression

Xiao Sun, Jiaxiang Shang, Shuang Liang, Yichen Wei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2602-2611

Regression based methods are not performing as well as detection based methods for human pose estimation. A central problem is that the structural information in the pose is not well exploited in the previous regression methods. In this work, we propose a structure-aware regression approach. It adopts a reparameterized pose representation using bones instead of joints. It exploits the joint connection structure to define a compositional loss function that encodes the long range interactions in the pose. It is simple, effective, and general for both 2D and 3D pose estimation in a unified setting. Comprehensive evaluation validates the effectiveness of our approach. It significantly advances the state-of-the-art on Human3.6M and is competitive with state-of-the-art results on MPII.

****

MUTAN: Multimodal Tucker Fusion for Visual Question Answering

Hedi Ben-younes, Remi Cadene, Matthieu Cord, Nicolas Thome; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2612-2620

Bilinear models provide an appealing framework for mixing and merging information in Visual Question Answering (VQA) tasks. They help to learn high level associations between question meaning and visual concepts in the image, but they suffer from huge dimensionality issues. We introduce MUTAN, a multimodal tensor-based Tucker decomposition to efficiently parametrize bilinear interactions between visual and textual representations. Additionally to the Tucker framework, we design a low-rank matrix-based decomposition to explicitly constrain the interaction rank. With MUTAN, we control the complexity of the merging scheme while keeping nice interpretable fusion relations. We show how the Tucker decomposition framework generalizes some of the latest VQA architectures, providing state-of-the-art results.

****

Revisiting IM2GPS in the Deep Learning Era

Nam Vo, Nathan Jacobs, James Hays; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2621-2630

Image geolocalization, inferring the geographic location of an image, is a challenging computer vision problem with many potential applications. The recent state-of-the-art approach to this problem is a deep image classification approach in which the world is spatially divided into bins and a deep network is trained to predict the correct bin for a given image. We propose to combine this approach with the original Im2GPS approach in which a query image is matched against a database of geotagged images and the location is inferred from the retrieved set. We estimate the geographic location of a query image by applying kernel density estimation to the locations of its nearest neighbors in the reference database. Interestingly, we find that the best features for our retrieval task are derived from networks trained with classification loss even though we do not use a classification approach at test time. Training with classification loss outperforms several deep feature learning methods (e.g. Siamese networks with contrastive of triplet loss) more typical for retrieval applications. Our simple approach achieves state-of-the-art geolocalization accuracy while also requiring significantly less training data.

****

Scene Parsing With Global Context Embedding

Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2631-2639

We present a scene parsing method that utilizes global context information based on both the parametric and non-parametric models. Compared to previous methods that only exploit the local relationship between objects, we train a context network based on scene similarities to generate feature representations for global contexts. In addition, these learned features are utilized to generate global and spatial priors for explicit classes inference. We then design modules to embed the feature representations and the priors into the segmentation network as additional global context cues. We show that the proposed method can eliminate false positives that are not compatible with the global context representations. Experiments on both the MIT ADE20K and PASCAL Context datasets show that the proposed method performs favorably against existing methods.

****

A Simple yet Effective Baseline for 3D Human Pose Estimation

Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2640-2649

Following the success of deep convolutional networks, state-of-the-art methods for 3d human pose estimation have focused on deep end-to-end systems that predict 3d joint locations given raw image pixels. Despite their excellent performance, it is often not easy to understand whether their remaining error stems from a limited 2d pose (visual) understanding, or from a failure to map 2d poses into 3-dimensional positions. With the goal of understanding these sources of error, we set out to build a system that given 2d joint locations predicts 3d positions. Much to our surprise, we have found that, with current technology, "lifting" ground truth 2d joint locations to 3d space is a task that can be solved with a remarkably low error rate: a relatively simple deep feed-forward network outperforms the best reported result by about 30% on Human3.6M, the largest publicly available 3d pose estimation benchmark. Furthermore, training our system on the output of an off-the-shelf state-of-the-art 2d detector (i.e., using images as input) yields state of the art results -- this includes an array of systems that have been trained end-to-end specifically for this task. Our results indicate that a large portion of the error of modern deep 3d pose estimation systems stems from their visual analysis, and suggests directions to further advance the state of the art in 3d human pose estimation.

****

Dual-Glance Model for Deciphering Social Relationships

Junnan Li, Yongkang Wong, Qi Zhao, Mohan S. Kankanhalli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2650-2659

Since the beginning of early civilizations, social relationships derived from each individual fundamentally form the basis of social structure in our daily life. In the computer vision literature, much progress has been made in scene understanding, such as object detection and scene parsing. Recent research focuses on the relationship between objects based on its functionality and geometrical relations. In this work, we aim to study the problem of social relationship recognition, in still images. We have proposed a dual-glance model for social relationship recognition, where the first glance fixates at the individual pair of interest and the second glance deploys attention mechanism to explore contextual cues. We have also collected a new large scale People in Social Context (PISC) dataset, which comprises of 22,670 images and 76,568 annotated samples from 9 types of social relationship. We provide benchmark results on the PISC dataset, and quali

tatively demonstrate the efficacy of the proposed model.
****

Sketching With Style: Visual Search With Sketches and Aesthetic Context
John Collomosse, Tu Bui, Michael J. Wilber, Chen Fang, Hailin Jin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2660-2668

We propose a novel measure of visual similarity for image retrieval that incorporates both structural and aesthetic (style) constraints. Our algorithm accepts a query as sketched shape, and a set of one or more contextual images specifying the desired visual aesthetic. A triplet network is used to learn a feature embedding capable of measuring style similarity independent of structure, delivering significant gains over previous networks for style discrimination. We incorporate this model within a hierarchical triplet network to unify and learn a joint space from two discriminatively trained streams for style and structure. We demonstrate that this space enables, for the first time, style-constrained sketch search over a diverse domain of digital artwork comprising graphics, paintings and drawings. We also briefly explore alternative query modalities.
****

Point Set Registration With Global-Local Correspondence and Transformation Estimation
Su Zhang, Yang Yang, Kun Yang, Yi Luo, Sim-Heng Ong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2669-2677

We present a new point set registration method with global-local correspondence and transformation estimation (GL-CATE). The geometric structures of point sets are exploited by combining the global feature, the point-to-point Euclidean distance, with the local feature, the shape distance (SD) which is based on the histograms generated by an elliptical Gaussian soft count strategy. By using a bi-directional deterministic annealing scheme to directly control the searching ranges of the two features, the mixture-feature Gaussian mixture model (MGMM) is constructed to recover the correspondences of point sets. A new vector based structure constraint term is formulated to regularize the transformation. The accuracy of transformation updating is improved by constraining spatial structure at both global and local scales. An annealing scheme is applied to progressively decrease the strength of the regularization and to achieve the maximum overlap. Both of the aforementioned processes are incorporated in the EM algorithm, an unified optimization framework. We test the performances of our GL-CATE in contour registration, sequence images, real images, medical images, fingerprint images and remote sensing images, and compare with eight state-of-the-art methods where our method shows favorable performances in most scenarios.
****

SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-Training on Indoor Segmentation?
John McCormac, Ankur Handa, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2678-2687

We introduce SceneNet RGB-D, a dataset providing pixel-perfect ground truth for scene understanding problems such as semantic segmentation, instance segmentation, and object detection. It also provides perfect camera poses and depth data, allowing investigation into geometric computer vision problems such as optical flow, camera pose estimation, and 3D scene labelling tasks. Random sampling permits virtually unlimited scene configurations, and here we provide 5M rendered RGB-D images from 16K randomly generated 3D trajectories in synthetic layouts, with random but physically simulated object configurations. We compare the semantic segmentation performance of network weights produced from pre-training on RGB images from our dataset against generic VGG-16 ImageNet weights. After fine-tuning on the SUN RGB-D and NYUv2 real-world datasets we find in both cases that the synthetically pre-trained network outperforms the VGG-16 weights. When synthetic pre-training includes a depth channel (something ImageNet cannot natively provide) the performance is greater still. This suggests that large-scale high-quality synthetic RGB datasets with task-specific labels can be more useful for pre-trai ning than real-world generic pre-training such as ImageNet. We host the dataset at http://robotvault.bitbucket.io/scenenet-rgbd.html

****

A Unified Model for Near and Remote Sensing

Scott Workman, Menghua Zhai, David J. Crandall, Nathan Jacobs; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2688-2697

We propose a novel convolutional neural network architecture for estimating geospatial functions such as population density, land cover, or land use. In our approach, we combine overhead and ground-level images in an end-to-end trainable neural network, which uses kernel regression and density estimation to convert features extracted from the ground-level images into a dense feature map. The output of this network is a dense estimate of the geospatial function in the form of a pixel-level labeling of the overhead image. To evaluate our approach, we created a large dataset of overhead and ground-level images from a major urban area with three sets of labels: land use, building function, and building age. We find that our approach is more accurate for all tasks, in some cases dramatically so.

****

Directionally Convolutional Networks for 3D Shape Segmentation

Haotian Xu, Ming Dong, Zichun Zhong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2698-2707

Previous approaches on 3D shape segmentation mostly rely on heuristic processing and hand-tuned geometric descriptors. In this paper, we propose a novel 3D shape representation learning approach, Directionally Convolutional Network (DCN), to solve the shape segmentation problem. DCN extends convolution operations from images to the surface mesh of 3D shapes. With DCN, we learn effective shape representations from raw geometric features, i.e., face normals and distances, to achieve robust segmentation. More specifically, a two-stream segmentation framework is proposed: one stream is made up by the proposed DCN with the face normals as the input, and the other stream is implemented by a neural network with the face distance histogram as the input. The learned shape representations from the two streams are fused by an element-wise product. Finally, Conditional Random Field (CRF) is applied to optimize the segmentation. Through extensive experiments conducted on benchmark datasets, we demonstrate that our approach outperforms the current state-of-the-arts (both classic and deep learning-based) on a large variety of 3D shapes.

****

AMAT: Medial Axis Transform for Natural Images

Stavros Tsogkas, Sven Dickinson; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2708-2717

We introduce Appearance-MAT (AMAT), a generalization of the medial axis transform for natural images, that is framed as a weighted geometric set cover problem. We make the following contributions: i) we extend previous medial point detection methods for color images, by associating each medial point with a local scale; ii) inspired by the invertibility property of the binary MAT, we also associate each medial point with a local encoding that allows us to invert the AMAT, reconstructing the input image; iii) we describe a clustering scheme that takes advantage of the additional scale and appearance information to group individual points into medial branches, providing a shape decomposition of the underlying image regions. In our experiments, we show state-of-the-art performance in medial point detection on Berkeley Medial AXes (BMAX500), a new dataset of medial axes based on the BSDS500 database, and good generalization on the SK506 and WH-SYMMAX datasets. We also measure the quality of reconstructed images from BMAX500, obtained by inverting their computed AMAT. Our approach delivers significantly better reconstruction quality wrt to three baselines, using just 10% of the image pixels. Our code and annotations are available at https://github.com/tsogkas/amat .

****

Deep Dual Learning for Semantic Image Segmentation

Ping Luo, Guangrun Wang, Liang Lin, Xiaogang Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2718-2726

Deep neural networks have advanced many computer vision tasks, because of their compelling capacities to learn from large amount of labeled data. However, their performances are not fully exploited in semantic image segmentation as the scale of training set is limited, where per-pixel labelmaps are expensive to obtain. To reduce labeling efforts, a natural solution is to collect additional images from Internet that are associated with image-level tags. Unlike existing works that treated labelmaps and tags as independent supervisions, we present a novel learning setting, namely dual image segmentation (DIS), which consists of two complementary learning problems that are jointly solved. One predicts labelmaps and tags from images, and the other reconstructs the images using the predicted labelmaps. DIS has three appealing properties. 1) Given an image with tags only, its labelmap can be inferred by leveraging the images and tags as constraints. The estimated labelmaps that capture accurate object classes and boundaries are used as ground truths in training to boost performance. 2) DIS is able to clean tags that have noises. 3) DIS significantly reduces the number of per-pixel annotations in training, while still achieves state-of-the-art performance. Extensive experiments demonstrate the effectiveness of DIS, which outperforms an existing best-performing baseline by 12.6% on Pascal VOC 2012 test set, without any post-processing such as CRF/MRF smoothing.
****

Regional Interactive Image Segmentation Networks
Jun Hao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, Jiashi Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2727-2735

The interactive image segmentation model allows users to iteratively add new inputs for refinement until a satisfactory result is finally obtained. Therefore, an ideal interactive segmentation model should learn to capture the user's intention with minimal interaction. However, existing models fail to fully utilize the valuable user input information in the segmentation refinement process and thus offer an unsatisfactory user experience. In order to fully exploit the user-provided information, we propose a new deep framework, called Regional Interactive Segmentation Network (RIS-Net), to expand the field-of-view of the given inputs to capture the local regional information surrounding them for local refinement. Additionally, RIS-Net adopts multiscale global contextual information to augment each local region for improving feature representation. We also introduce click discount factors to develop a novel optimization strategy for more effective end-to-end training. Comprehensive evaluations on four challenging datasets well demonstrate the superiority of the proposed RIS-Net over other state-of-the-art approaches.
****

Learning Efficient Convolutional Networks Through Network Slimming
Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, Changshui Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2736-2744

The deployment of deep convolutional neural networks (CNNs) in many real world applications is largely hindered by their high computational cost. In this paper, we propose a novel learning scheme for CNNs to simultaneously 1) reduce the model size; 2) decrease the run-time memory footprint; and 3) lower the number of computing operations, without compromising accuracy. This is achieved by enforcing channel-level sparsity in the network in a simple but effective way. Different from many existing approaches, the proposed method directly applies to modern CNN architectures, introduces minimum overhead to the training process, and requires no special software/hardware accelerators for the resulting models. We call our approach network slimming, which takes wide and large networks as input models, but during training insignificant channels are automatically identified and pruned afterwards, yielding thin and compact models with comparable accuracy. We empirically demonstrate the effectiveness of our approach with several state-of-the-art CNN models, including VGGNet, ResNet and DenseNet, on various image classification datasets. For VGGNet, a multi-pass version of network slimming gives a 20x reduction in model size and a 5x reduction in computing operations.
****

CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training
Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, Gang Hua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2745-2754

We present variational generative adversarial networks, a general learning framework that combines a variational auto-encoder with a generative adversarial network, for synthesizing images in fine-grained categories, such as faces of a specific person or objects in a category. Our approach models an image as a composition of label and latent attributes in a probabilistic model. By varying the fine-grained category label fed into the resulting generative model, we can generate images in a specific category with randomly drawn values on a latent attribute vector. Our approach has two novel aspects. First, we adopt a cross entropy loss for the discriminative and classifier network, but a mean discrepancy objective for the generative network. This kind of asymmetric loss function makes the GAN training more stable. Second, we adopt an encoder network to learn the relationship between the latent space and the real image space, and use pairwise feature matching to keep the structure of generated images. We experiment with natural images of faces, flowers, and birds, and demonstrate that the proposed models are capable of generating realistic and diverse samples with fine-grained category labels. We further show that our models can be applied to other tasks, such as image inpainting, super-resolution, and data augmentation for training better face recognition models.

****

Universal Adversarial Perturbations Against Semantic Image Segmentation
Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, Volker Fischer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2755-2764

While deep learning is remarkably successful on perceptual tasks, it was also shown to be vulnerable to adversarial perturbations of the input. These perturbations denote noise added to the input that was generated specifically to fool the system while being quasi-imperceptible for humans. More severely, there even exist universal perturbations that are input-agnostic but fool the network on the majority of inputs. While recent work has focused on image classification, this work proposes attacks against semantic image segmentation: we present an approach for generating (universal) adversarial perturbations that make the network yield a desired target segmentation as output. We show empirically that there exist barely perceptible universal noise patterns which result in nearly the same predicted segmentation for arbitrary inputs. Furthermore, we also show the existence of universal noise which removes a target class (e.g., all pedestrians) from the segmentation while leaving the segmentation mostly unchanged otherwise.

****

Associative Domain Adaptation
Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2765-2773

We propose "associative domain adaptation", a novel technique for end-to-end domain adaptation with neural networks, the task of inferring class labels for an unlabeled target domain based on the statistical properties of a labeled source domain. Our training scheme follows the paradigm that in order to effectively derive class labels for the target domain, a network should produce statistically domain invariant embeddings, while minimizing the classification error on the labeled source domain. We accomplish this by reinforcing "associations" between source and target data directly in embedding space. Our method can easily be added to any existing classification network with no structural and almost no computational overhead. We demonstrate the effectiveness of our approach on various benchmarks and achieve state-of-the-art results across the board with a generic convolutional neural network architecture not specifically tuned to the respective tasks. Finally, we show that the proposed association loss produces embeddings that are more effective for domain adaptation compared to methods employing maximum mean discrepancy as a similarity measure in embedding space.

****

Introspective Neural Networks for Generative Modeling

Justin Lazarow, Long Jin, Zhuowen Tu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2774-2783

We study unsupervised learning by developing a generative model built from progressively learned deep convolutional neural networks. The resulting generator is additionally a discriminator, capable of "introspection" in a sense --- being able to self-evaluate the difference between its generated samples and the given training data. Through repeated discriminative learning, desirable properties of modern discriminative classifiers are directly inherited by the generator. Specifically, our model learns a sequence of CNN classifiers using a synthesis-by-classification algorithm. In the experiments, we observe encouraging results on a number of applications including texture modeling, artistic style transferring, face modeling, and unsupervised feature learning.

****

Towards a Unified Compositional Model for Visual Pattern Modeling

Wei Tang, Pei Yu, Jiahuan Zhou, Ying Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2784-2793

Compositional models represent visual patterns as hierarchies of meaningful and reusable parts. They are attractive to vision modeling due to their ability to decompose complex patterns into simpler ones and resolve the low-level ambiguities in high-level image interpretations. However, current compositional models separate structure and part discovery from parameter estimation, which generally leads to suboptimal learning and fitting of the model. Moreover, the commonly adopted latent structural learning is not scalable for deep architectures. To address these difficult issues for compositional models, this paper quests for a unified framework for compositional pattern modeling, inference and learning. Represented by And-Or graphs (AOGs), it jointly models the compositional structure, parts, features, and composition/sub-configuration relationships. We show that the inference algorithm of the proposed framework is equivalent to a feed-forward network. Thus, all the parameters can be learned efficiently via the highly-scalable back-propagation (BP) in an end-to-end fashion. We validate the model via the task of handwritten digit recognition. By visualizing the processes of bottom-up composition and top-down parsing, we show that our model is fully interpretable, being able to learn the hierarchical compositions from visual primitives to visual patterns at increasingly higher levels. We apply this new compositional model to natural scene character recognition and generic object detection. Experimental results have demonstrated its effectiveness.

****

Least Squares Generative Adversarial Networks

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, Stephen Paul Smolley; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2794-2802

Unsupervised learning with generative adversarial networks (GANs) has proven hugely successful. Regular GANs hypothesize the discriminator as a classifier with the sigmoid cross entropy loss function. However, we found that this loss function may lead to the vanishing gradients problem during the learning process. To overcome such a problem, we propose in this paper the Least Squares Generative Adversarial Networks (LSGANs) which adopt the least squares loss function for the discriminator. We show that minimizing the objective function of LSGAN yields minimizing the Pearson Chi^2 divergence. There are two benefits of LSGANs over regular GANs. First, LSGANs are able to generate higher quality images than regular GANs. Second, LSGANs perform more stable during the learning process. We evaluate LSGANs on LSUN and CIFAR-10 datasets and the experimental results show that the images generated by LSGANs are of better quality than the ones generated by regular GANs. We also conduct two comparison experiments between LSGANs and regular GANs to illustrate the stability of LSGANs.

****

Centered Weight Normalization in Accelerating Training of Deep Neural Networks

Lei Huang, Xianglong Liu, Yang Liu, Bo Lang, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2803-2811

Training deep neural networks is difficult for the pathological curvature problem. Re-parameterization is an effective way to relieve the problem by learning the curvature approximately or constraining the solutions of weights with good properties for optimization. This paper proposes to re-parameterize the input weight of each neuron in deep neural networks by normalizing it with zero-mean and unit-norm, followed by a learnable scalar parameter to adjust the norm of the weight. This technique effectively stabilizes the distribution implicitly. Besides, it improves the conditioning of the optimization problem and thus accelerates the training of deep neural networks. It can be wrapped as a linear module in practice and plugged in any architecture to replace the standard linear module. We highlight the benefits of our method on both multi-layer perceptrons and convolutional neural networks, and demonstrate its scalability and efficiency on SVHN, CIFAR-10, CIFAR-100 and ImageNet datasets.

****

Deep Growing Learning

Guangcong Wang, Xiaohua Xie, Jianhuang Lai, Jiaxuan Zhuo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2812-2820

Semi-supervised learning (SSL) is an import paradigm to make full use of a large amount of unlabeled data in machine learning. A bottleneck of SSL is the overfitting problem when training over the limited labeled data, especially on a complex model like a deep neural network. To get around this bottleneck, we propose a bio-inspired SSL framework on deep neural network, namely Deep Growing Learning (DGL). Specifically, we formulate the SSL as an EM-like process, where the deep network alternately iterates between automatically growing convolutional layers and selecting reliable pseudo-labeled data for training. The DGL guarantees that a shallow neural network is trained with labeled data, while a deeper neural network is trained with growing amount of reliable pseudo-labeled data, so as to alleviate the overfitting problem. Experiments on different visual recognition tasks have verified the effectiveness of DGL.

****

Smart Mining for Deep Metric Learning

Ben Harwood, Vijay Kumar B G, Gustavo Carneiro, Ian Reid, Tom Drummond; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2821-2829

To solve deep metric learning problems and produce feature embeddings, current methodologies will commonly use a triplet model to minimise the relative distance between samples from the same class and maximise the relative distance between samples from different classes. Though successful, the training convergence of this triplet model can be compromised by the fact that the vast majority of the training samples will produce gradients with magnitudes that are close to zero. This issue has motivated the development of methods that explore the global structure of the embedding and other methods that explore hard negative/positive mining. The effectiveness of such mining methods is often associated with intractable computational requirements. In this paper, we propose a novel deep metric learning method that combines the triplet model and the global structure of the embedding space. We rely on a smart mining procedure that produces effective training samples for a low computational cost. In addition, we propose an adaptive controller that automatically adjusts the smart mining hyper-parameters and speeds up the convergence of the training process. We show empirically that our proposed method allows for fast and more accurate training of triplet ConvNets than other competing mining methods. Additionally, we show that our method achieves new state-of-the-art embedding results for CUB-200-2011 and Cars196 datasets.

****

Temporal Generative Adversarial Nets With Singular Value Clipping

Masaki Saito, Eiichi Matsumoto, Shunta Saito; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2830-2839

In this paper, we propose a generative model, Temporal Generative Adversarial Nets (TGAN), which can learn a semantic representation of unlabeled videos, and is capable of generating videos. Unlike existing Generative Adversarial Nets (GAN)-based methods that generate videos with a single generator consisting of 3D dec onvolutional layers, our model exploits two different types of generators: a temporal generator and an image generator. The temporal generator takes a single latent variable as input and outputs a set of latent variables, each of which corresponds to an image frame in a video. The image generator transforms a set of such latent variables into a video. To deal with instability in training of GAN with such advanced networks, we adopt a recently proposed model, Wasserstein GAN, and propose a novel method to train it stably in an end-to-end manner. The experimental results demonstrate the effectiveness of our methods.

****

Sampling Matters in Deep Embedding Learning

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, Philipp Krahenbuhl; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2840-2848

Deep embeddings answer one simple question: How similar are two images? Learning these embeddings is the bedrock of verification, zero-shot learning, and visual search. The most prominent approaches optimize a deep convolutional network with a suitable loss function, such as contrastive loss or triplet loss. While a rich line of work focuses solely on the loss functions, we show in this paper that selecting training examples plays an equally important role. We propose distance weighted sampling, which selects more informative and stable examples than traditional approaches. In addition, we show that a simple margin based loss is sufficient to outperform all other loss functions. We evaluate our approach on the CUB200-2011, CAR196, and the Stanford Online Products datasets for image retrieval and clustering, and on the LFW dataset for face verification. Our method achieves state-of-the-art performance on all of them.

****

DualGAN: Unsupervised Dual Learning for Image-To-Image Translation

Zili Yi, Hao Zhang, Ping Tan, Minglun Gong; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2849-2857

Conditional Generative Adversarial Networks (GANs) for cross-domain image-to-image translation have made much progress recently. Depending on the task complexity, thousands to millions of labeled image pairs are needed to train a conditional GAN. However, human labeling is expensive, even impractical, and large quantities of data may not always be available. Inspired by dual learning from natural language translation, we develop a novel mechanism, which enables image translators to be trained from two sets of images from two domains. In our architecture, the primal GAN learns to translate images from domain U to those in domain V, while the dual GAN learns to invert the task. The closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed. Hence a loss function that accounts for the reconstruction error of images can be used to train the translators. Experiments on multiple image translation tasks with unlabeled data show considerable performance gain of DualGAN over a single GAN. For some tasks, DualGAN can even achieve comparable or slightly better results than conditional GAN trained on fully labeled data.

****

Learning View-Invariant Features for Person Identification in Temporally Synchronized Videos Taken by Wearable Cameras

Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, Hongkai Yu, Dazhou Guo, Song Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2858-2866

In this paper, we study the problem of Cross-View Person Identification (CVPI), which aims at identifying the same person from temporally synchronized videos taken by different wearable cameras. Our basic idea is to utilize the human motion consistency for CVPI, where human motion can be computed by optical flow. However, optical flow is view-variant -- the same person's optical flow in different videos can be very different due to view angle change. In this paper, we attempt to utilize 3D human-skeleton sequences to learn a model that can extract view-invariant motion features from optical flows in different views. For this purpose, we use 3D Mocap database to build a synthetic optical flow dataset and train a Triplet Network (TN) consisting of three sub-networks: two for optical flow seq

uences from different views and one for the underlying 3D Mocap skeleton sequence. Finally, sub-networks for optical flows are used to extract view-invariant features for CVPI. Experimental results show that, using only the motion information, the proposed method can achieve comparable performance with the state-of-the-art methods. Further combination of the proposed method with an appearance-based method achieves new state-of-the-art performance.
****

MarioQA: Answering Questions by Watching Gameplay Videos

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, Bohyung Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2867-2875

We present a framework to analyze various aspects of models for video question answering (VideoQA) using customizable synthetic datasets, which are constructed automatically from gameplay videos. Our work is motivated by the fact that existing models are often tested only on datasets that require excessively high-level reasoning or mostly contain instances accessible through single frame inferences. Hence, it is difficult to measure capacity and flexibility of trained models, and existing techniques often rely on ad-hoc implementations of deep neural networks without clear insight into datasets and models. We are particularly interested in understanding temporal relationships between video events to solve VideoQA problems; this is because reasoning temporal dependency is one of the most distinct components in videos from images. To address this objective, we automatically generate a customized synthetic VideoQA dataset using Super Mario Bros. gameplay videos so that it contains events with different levels of reasoning complexity. Using the dataset, we show that properly constructed datasets with events in various complexity levels are critical to learn effective models and improve overall performance.
****

SBGAR: Semantics Based Group Activity Recognition

Xin Li, Mooi Choo Chuah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2876-2885

Activity recognition has become an important function in many emerging computer vision applications e.g. automatic video surveillance system, human-computer interaction application, and video recommendation system, etc. In this paper, we propose a novel semantics based group activity recognition scheme, namely SBGAR, which achieves higher accuracy and efficiency than existing group activity recognition methods. SBGAR consists of two stages: in stage I, we use a LSTM model to generate a caption for each video frame; in stage II, another LSTM model is trained to predict the final activity categories based on these generated captions. We evaluate SBGAR using two well-known datasets: the Collective Activity Dataset and the Volleyball Dataset. Our experimental results show that SBGAR improves the group activity recognition accuracy with shorter computation time compared to the state-of-the-art methods.
****

Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video

Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, Dima Damen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2886-2894

Manual annotations of temporal bounds for object interactions (i.e. start and end times) are typical training input to recognition, localization and detection algorithms. For three publicly available egocentric datasets, we uncover inconsistencies in ground truth temporal bounds within and across annotators and datasets. We systematically assess the robustness of state-of-the-art approaches to changes in labeled temporal bounds, for object interaction recognition. As boundaries are trespassed, a drop of up to 10% is observed for both Improved Dense Trajectories and Two-Stream Convolutional Neural Network. We demonstrate that such disagreement stems from a limited understanding of the distinct phases of an action, and propose annotating based on the Rubicon Boundaries, inspired by a similarly named cognitive model, for consistent temporal bounds of object interactions. Evaluated on a public dataset, we report a 4% increase in overall accuracy, and an increase in accuracy for 55% of classes when Rubicon Boundaries are used for temporal annotations.

****

Unmasking the Abnormal Events in Video

Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, Marius Popescu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2895-2903

We propose a novel framework for abnormal event detection in video that requires no training sequences. Our framework is based on unmasking, a technique previously used for authorship verification in text documents, which we adapt to our task. We iteratively train a binary classifier to distinguish between two consecutive video sequences while removing at each step the most discriminant features. Higher training accuracy rates of the intermediately obtained classifiers represent abnormal events. To the best of our knowledge, this is the first work to apply unmasking for a computer vision task. We compare our method with several state-of-the-art supervised and unsupervised methods on four benchmark data sets. The empirical results indicate that our abnormal event detection framework can achieve state-of-the-art results, while running in real-time at 20 frames per second.

****

Chained Multi-Stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection

Mohammadreza Zolfaghari, Gabriel L. Oliveira, Nima Sedaghat, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2904-2913

General human action recognition requires understanding of various visual cues. In this paper, we propose a network architecture that computes and integrates the most important visual cues for action recognition: pose, motion, and the raw images. For the integration, we introduce a Markov chain model which adds cues successively. The resulting approach is efficient and applicable to action classification as well as to spatial and temporal action localization. The two contributions clearly improve the performance over respective baselines. The overall approach achieves state-of-the-art action classification performance on HMDB51, J-HMDB and NTU RGB+D datasets. Moreover, it yields state-of-the-art spatio-temporal action localization results on UCF101 and J-HMDB.

****

Temporal Action Detection With Structured Segment Networks

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, Dahua Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2914-2923

Detecting actions in untrimmed videos is an important yet challenging task. In this paper, we present the structured segment network (SSN), a novel framework which models the temporal structure of each action instance via a structured temporal pyramid. On top of the pyramid, we further introduce a decomposed discriminative model comprising two classifiers, respectively for classifying actions and determining completeness. This allows the framework to effectively distinguish positive proposals from background or incomplete ones, thus leading to both accurate recognition and localization. These components are integrated into a unified network that can be efficiently trained in an end-to-end fashion. Additionally, a simple yet effective temporal action proposal scheme, dubbed temporal actionness grouping (TAG) is devised to generate high quality action proposals. On two challenging benchmarks, THUMOS'14 and ActivityNet, our method remarkably outperforms previous state-of-the-art methods, demonstrating superior accuracy and strong adaptivity in handling actions with various temporal structures.

****

Jointly Recognizing Object Fluents and Tasks in Egocentric Videos

Yang Liu, Ping Wei, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2924-2932

This paper addresses the problem of jointly recognizing object fluents and tasks in egocentric videos. Fluents are the changeable attributes of objects. Tasks a

re goal-oriented human activities which interact with objects and aim to change some attributes of the objects. The process of executing a task is a process to change the object fluents over time. We propose a hierarchical model to represent tasks as concurrent and sequential object fluents. In a task, different fluents closely interact with each other both in spatial and temporal domains. Given an egocentric video, a beam search algorithm is applied to jointly recognizing the object fluents in each frame, and the task of the entire video. We collected a large scale egocentric video dataset of tasks and fluents. This dataset contains 14 categories of tasks, 25 object classes, 21 categories of object fluents, 809 video sequences, and approximately 333,000 video frames. The experimental results on this dataset prove the strength of our method.
****
Transferring Objects: Joint Inference of Container and Human Pose
Hanqing Wang, Wei Liang, Lap-Fai Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2933-2941
Transferring objects from one place to another place is a common task performed by human in daily life. During this process, it is usually intuitive for humans to choose an object as a proper container and to use an efficient pose to carry objects; yet, it is non-trivial for current computer vision and machine learning algorithms. In this paper, we propose an approach to jointly infer container and human pose for transferring objects by minimizing the costs associated both object and pose candidates. Our approach predicts which object to choose as a container while reasoning about how humans interact with physical surroundings to accomplish the task of transferring objects given visual input. In the learning phase, the presented method learns how humans make rational choices of containers and poses for transferring different objects, as well as the physical quantities required by the transfer task (e.g., compatibility between container and containee, energy cost of carrying pose) via a structured learning approach. In the inference phase, given a scanned 3D scene with different object candidates and a dictionary of human poses, our approach infers the best object as a container together with human pose for transferring a given object.
****
Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention
Jinkyu Kim, John Canny; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2942-2950
Deep neural perception and control networks are likely to be a key component of self-driving vehicles. These models need to be explainable - they should provide easy-to-interpret rationales for their behavior - so that passengers, insurance companies, law enforcement, developers etc., can understand what triggered a particular behavior. Here we explore the use of visual explanations. These explanations take the form of real-time highlighted regions of an image that causally influence the network's output (steering control). Our approach is two-stage. In the first stage, we use a visual attention model to train a convolution network end-to-end from images to steering angle. The attention model highlights image regions that potentially influence the network's output. Some of these are true influences, but some are spurious. We then apply a causal filtering step to determine which input regions actually influence the output. This produces more succinct visual explanations and more accurately exposes the network's behavior. We demonstrate the effectiveness of our model on three datasets totaling 16 hours of driving. We first show that training with attention does not degrade the performance of the end-to-end network. Then we show that the network causally cues on a variety of features that are used by humans while driving.
****
Learning Cooperative Visual Dialog Agents With Deep Reinforcement Learning
Abhishek Das, Satwik Kottur, Jose M. F. Moura, Stefan Lee, Dhruv Batra; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2951-2960
We introduce the first goal-driven training for visual question answering and dialog agents. Specifically, we pose a cooperative `image guessing' game between two agents -- Qbot and Abot -- who communicate in natural language dialog so that

Qbot can select an unseen image from a lineup of images. We use deep reinforcement learning (RL) to end-to-end learn the policies of these agents -- from pixels to multi-agent multi-round dialog to game reward. We demonstrate two experimental results. First, as a `sanity check' demonstration of pure RL (from scratch), we show results on a synthetic world, where the agents communicate in ungrounded vocabulary, ie, symbols with no pre-specified meanings (X, Y, Z). We find that two bots invent their own communication protocol and start using certain symbols to ask/answer about certain visual attributes (shape/color/size). Thus, we demonstrate the emergence of grounded language and communication among `visual' dialog agents with no human supervision at all. Second, we conduct large-scale real-image experiments on the VisDial dataset, where we pretrain on dialog data and show that the RL fine-tuned agents significantly outperform supervised pretraining. Interestingly, the RL Qbot learns to ask questions that Abot is good at, ultimately resulting in more informative dialog and a better team.

****

Mask R-CNN

Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969

We present a conceptually simple, flexible, and general framework for object instance segmentation. Our approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps. Moreover, Mask R-CNN is easy to generalize to other tasks, e.g., allowing us to estimate human poses in the same framework. We show top results in all three tracks of the COCO suite of challenges, including instance segmentation, bounding-box object detection, and person keypoint detection. Without tricks, Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners. We hope our simple and effective approach will serve as a solid baseline and help ease future research in instance-level recognition. Code will be made available.

****

Towards Diverse and Natural Image Descriptions via a Conditional GAN

Bo Dai, Sanja Fidler, Raquel Urtasun, Dahua Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2970-2979

Despite the substantial progress in recent years, the problem of image captioning remains far from being satisfactorily tackled. Sentences produced by existing methods, e.g. those based on LSTM, are often overly rigid and lacking in variability. This issue is related to a learning principle widely used in practice, that is, to maximize the likelihood of training samples. This principle encourages the high resemblance to the "ground-truths", while suppressing other reasonable expressions. Conventional evaluation metrics, e.g. BLEU and METEOR, also favor such restrictive methods. In this paper, we explore an alternative approach, with an aim to improve the naturalness and diversity - two essential properties of human expressions. Specifically, we propose a new framework based on Conditional Generative Adversarial Networks (CGAN), which jointly learns a generator to produce descriptions conditioned on images and an evaluator to assess how well a description fits the visual content. It is noteworthy that training a sequence generator is nontrivial. We overcome the difficulty by Policy Gradient, a strategy stemming from Reinforcement Learning, which allows the generator to receive early feedbacks along the way. We tested our method on two large datasets, where it performed competitively against real people in our user study and outperformed other methods on various tasks.

****

Focal Loss for Dense Object Detection

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988

The highest accuracy object detectors to date are based on a two-stage approach popularized by R-CNN, where a classifier is applied to a sparse set of candidate object locations. In contrast, one-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far. In this paper, we investigate why this is the case. We discover that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. We propose to address this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. Our novel Focal Loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. To evaluate the effectiveness of our loss, we design and train a simple dense detector we call RetinaNet. Our results show that when trained with the focal loss, RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy of all existing state-of-the-art two-stage detectors.

****

Inferring and Executing Programs for Visual Reasoning

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2989-2998

Existing methods for visual reasoning attempt to directly map inputs to outputs using black-box architectures without explicitly modeling the underlying reasoning processes. As a result, these black-box models often learn to exploit biases in the data rather than learning to perform visual reasoning. Inspired by module networks, this paper proposes a model for visual reasoning that consists of a program generator that constructs an explicit representation of the reasoning process to be performed, and an execution engine that executes the resulting program to produce an answer. Both the program generator and the execution engine are implemented by neural networks, and are trained using a combination of backpropagation and REINFORCE. Using the CLEVR benchmark for visual reasoning, we show that our model significantly outperforms strong baselines and generalizes better in a variety of settings.

****

Visual Forecasting by Imitating Dynamics in Natural Sequences

Kuo-Hao Zeng, William B. Shen, De-An Huang, Min Sun, Juan Carlos Niebles; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999-3008

We introduce a general framework for visual forecasting, which directly imitates visual sequences without additional supervision. As a result, our model can be applied at several semantic levels and does not require any domain knowledge or handcrafted features. We achieve this by formulating visual forecasting as an inverse reinforcement learning (IRL) problem, and directly imitate the dynamics in natural sequences from their raw pixel values. The key challenge is the high-dimensional and continuous state-action space that prohibits the application of previous IRL algorithms. We address this computational bottleneck by extending recent progress in model-free imitation with trainable deep feature representations, which (1) bypasses the exhaustive state-action pair visits in dynamic programming by using a dual formulation and (2) avoids explicit state sampling at gradient computation using a deep feature reparametrization. This allows us to apply IRL at scale and directly imitate the dynamics in high-dimensional continuous visual sequences from the raw pixel values. We evaluate our approach at three different level-of-abstraction, from low level pixels to higher level semantics: future frame generation, action anticipation, visual story forecasting. At all levels, our approach outperforms existing methods.

****

TorontoCity: Seeing the World With a Million Eyes

Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3009-3017

In this paper we introduce the TorontoCity benchmark, which covers the full grea ter Toronto area (GTA) with 712.5km2 of land, 8439km of road and around 400, 000 buildings. Our benchmark provides different perspectives of the world captured from airplanes, drones and cars driving around the city. Manually labeling such a large scale dataset is infeasible. Instead, we propose to utilize different sources of high-precision maps to create our ground truth. Towards this goal, we develop algorithms that allow us to align all data sources with the maps while requiring minimal human supervision. We have designed a wide variety of tasks including building height estimation (reconstruction), road centerline and curb extraction, building instance segmentation, building contour extraction (reorganization), semantic labeling and scene type classification (recognition). Our pilot study shows that most of these tasks are still difficult for modern convolutional neural networks.

****

Low-Shot Visual Recognition by Shrinking and Hallucinating Features

Bharath Hariharan, Ross Girshick; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3018-3027

Low-shot visual learning--the ability to recognize novel object categories from very few examples--is a hallmark of human visual intelligence. Existing machine learning approaches fail to generalize in the same way. To make progress on this foundational problem, we present a low- shot learning benchmark on complex images that mimics challenges faced by recognition systems in the wild. We then propose (1) representation regularization techniques, and (2) techniques to hallucinate additional training examples for data-starved classes. Together, our methods improve the effectiveness of convolutional networks in low-shot learning, improving the one-shot accuracy on novel classes by 2.3x on the challenging ImageNet dataset.

****

A Coarse-Fine Network for Keypoint Localization

Shaoli Huang, Mingming Gong, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3028-3037

We propose a coarse-fine network (CFN) that exploits multi-level supervisions for keypoint localization. Recently, convolutional neural networks (CNNs)-based methods have achieved great success due to the powerful hierarchical features in CNNs. These methods typically use confidence maps generated from ground-truth keypoint locations as supervisory signals. However, while some keypoints can be easily located with high accuracy, many of them are hard to localize due to appearance ambiguity. Thus, using strict supervision often fails to detect keypoints that are difficult to locate accurately. To target this problem, we develop a keypoint localization network composed of several coarse detector branches, each of which is built on top of a feature layer in a CNN, and a fine detector branch built on top of multiple feature layers. We supervise each branch by a specified label map to explicate a certain supervision strictness level. All the branches are unified principally to produce the final accurate keypoint locations. We demonstrate the efficacy, efficiency, and generality of our method on several benchmarks for multiple tasks including bird part localization and human body pose estimation. Especially, our method achieves 72.2% AP on the 2016 COCO Keypoints Challenge dataset, which is an 18% improvement over the winning entry.

****

Detect to Track and Track to Detect

Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3038-3046

Recent approaches for high accuracy detection and tracking of object categories in video consist of complex multistage solutions that become more cumbersome each year. In this paper we propose a ConvNet architecture that jointly performs detection and tracking, solving the task in a simple and effective way. Our contributions are threefold: (i) we set up a ConvNet architecture for simultaneous detection and tracking, using a multi-task objective for frame-based object detection and across-frame track regression; (ii) we introduce correlation features that represent object co-occurrences across time to aid the ConvNet during tracking ; and (iii) we link the frame level detections based on our across-frame trackle ts to produce high accuracy detections at the video level. Our ConvNet architecture for spatiotemporal object detection is evaluated on the large-scale ImageNet VID dataset where it achieves state-of-the-art results. Our approach provides better single model performance than the winning method of the last ImageNet challenge while being conceptually much simpler. Finally, we show that by increasing the temporal stride we can dramatically increase the tracker speed.

****

Single Shot Text Detector With Regional Attention

Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, Xiaolin Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3047-3055

We present a novel single-shot text detector that directly outputs word-level bounding boxes in a natural image. We propose an attention mechanism which roughly identifies text regions via an automatically learned attentional map. This substantially suppresses background interference in the convolutional features, which is the key to producing accurate inference of words, particularly at extremely small sizes. This results in a single model that essentially works in a coarse-to-fine manner. It departs from recent FCN-based text detectors which cascade multiple FCN models to achieve an accurate prediction. Furthermore, we develop a hierarchical inception module which efficiently aggregates multi-scale inception features. This enhances local details, and also encodes strong context information, allowing the detector to work reliably on multi-scale and multi-orientation text with single-scale images. Our text detector achieves an F-measure of 77% on the ICDAR 2015 benchmark, advancing the state-of-the-art results.

****

SubUNets: End-To-End Hand Shape and Continuous Sign Language Recognition

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Richard Bowden; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3056-3065

We propose a novel deep learning approach to solve simultaneous alignment and recognition problems (referred to as "Sequence-to-sequence" learning). We decompose the problem into a series of specialised expert systems referred to as SubUNets. The spatio-temporal relationships between these SubUNets are then modelled to solve the task, while remaining trainable end-to-end. The approach mimics human learning and educational techniques, and has a number of significant advantages. SubUNets allow us to inject domain-specific expert knowledge into the system regarding suitable intermediate representations. They also allow us to implicitly perform transfer learning between different interrelated tasks, which also allows us to exploit a wider range of more varied data sources. In our experiments we demonstrate that each of these properties serves to significantly improve the performance of the overarching recognition system, by better constraining the learning problem. The proposed techniques are demonstrated in the challenging domain of sign language recognition. We demonstrate state-of-the-art performance on hand-shape recognition outperforming previous techniques by more than 30%). Furthermore, we are able to obtain comparable sign recognition rates to previous research, without the need for an alignment step to segment out the signs for recognition.

****

A Spatiotemporal Oriented Energy Network for Dynamic Texture Recognition

Isma Hadji, Richard P. Wildes; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3066-3074

This paper presents a novel hierarchical spatiotemporal orientation representation for spacetime image analysis. It is designed to combine the benefits of the multilayer architecture of ConvNets and a more controlled approach to spacetime analysis. A distinguishing aspect of the approach is that unlike most contemporary convolutional networks no learning is involved; rather, all design decisions are specified analytically with theoretical motivations. This approach makes it possible to understand what information is being extracted at each stage and layer of processing as well as to minimize heuristic choices in design. Another key aspect of the network is its recurrent nature, whereby the output of each layer of processing feeds back to the input. To keep the network size manageable acros s layers, a novel cross-channel feature pooling is proposed. The multilayer architecture that results systematically reveals hierarchical image structure in terms of multiscale, multiorientation properties of visual spacetime. To illustrate its utility, the network has been applied to the task of dynamic texture recognition. Empirical evaluation on multiple standard datasets shows that it sets a new state-of-the-art.

****

Probabilistic Structure From Motion With Objects (PSfMO)
Paul Gay, Cosimo Rubino, Vaibhav Bansal, Alessio Del Bue; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3075-3084

In this paper we deal with the problem of recovering affine camera calibration and objects position/occupancy from multi-view images using the information from image detections. We show that remarkable object localisation and volumetric occupancy can be recovered by including both geometrical constraints and prior information given by objects CAD models from the ShapeNet dataset. This can be done by recasting the problem in the context of a probabilistic framework based on Probabilistic PCA that includes both the object semantic priors together with the multi-view geometrical constraints. We present results on synthetic and real datasets to show the validity of our approach and improvements with respect to previous approaches. In particular, the statistical priors are key to obtain reliable 3D reconstruction especially when the input detections are noisy, a likely case in real scenarios.

****

A 3D Morphable Model of Craniofacial Shape and Texture Variation
Hang Dai, Nick Pears, William A. P. Smith, Christian Duncan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3085-3093

We present a fully automatic pipeline to train 3D Morphable Models (3DMMs), with contributions in pose normalisation, dense correspondence using both shape and texture information, and high quality, high resolution texture mapping. We propose a dense correspondence system, combining a hierarchical parts-based template morphing framework in the shape channel and a refining optical flow in the texture channel. The texture map is generated using raw texture images from five views. We employ a pixel-embedding method to maintain the texture map at the same high resolution as the raw texture images, rather than using per-vertex color maps. The high quality texture map is then used for statistical texture modelling. The Headspace dataset used for training includes demographic information about each subject, allowing for the construction of both global 3DMMs and models tailored for specific gender and age groups. We build both global craniofacial 3DMMs and demographic sub-population 3DMMs from more than 1200 distinct identities. To our knowledge, we present the first public 3DMM of the full human head in both shape and texture: the Liverpool-York Head Model. Furthermore, we analyse the 3DMMs in terms of a range of performance metrics. Our evaluations reveal that the training pipeline constructs state-of-the-art models.

****

Multi-View Dynamic Shape Refinement Using Local Temporal Integration
Vincent Leroy, Jean-Sebastien Franco, Edmond Boyer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3094-3103

We consider 4D shape reconstructions in multi-view environments and investigate how to exploit temporal redundancy for precision refinement. In addition to being beneficial to many dynamic multi-view scenarios this also enables larger scenes where such increased precision can compensate for the reduced spatial resolution per image frame. With precision and scalability in mind, we propose a symmetric (non-causal) local time-window geometric integration scheme over temporal sequences, where shape reconstructions are refined framewise by warping local and reliable geometric regions of neighboring frames to them. This is in contrast to recent comparable approaches targeting a different context with more compact scenes and real-time applications. These usually use a single dense volumetric update space or geometric template, which they causally track and update globally frame by frame, with limitations in scalability for larger scenes and in topology and precision with a template based strategy. Our template less and local approa ch is a first step towards temporal shape super-resolution. We show that it improves reconstruction accuracy by considering multiple frames. To this purpose, and in addition to real data examples, we introduce a multi-camera synthetic dataset that provides ground-truth data for mid-scale dynamic scenes.

****

Learning Hand Articulations by Hallucinating Heat Distribution

Chiho Choi, Sangpil Kim, Karthik Ramani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3104-3113

We propose a robust hand pose estimation method by learning hand articulations from depth features and auxiliary modality features. As an additional modality to depth data, we present a function of geometric properties on the surface of the hand described by heat diffusion. The proposed heat distribution descriptor is robust to identify the keypoints on the surface as it incorporates both the local geometry of the hand and global structural representation at multiple time scales. Along this line, we train our heat distribution network to learn the geometrically descriptive representations from the proposed descriptors with the fingertip position labels. Then the hallucination network is guided to mimic the intermediate responses of the heat distribution modality from a paired depth image. We use the resulting geometrically informed responses together with the discriminative depth features estimated from the depth network to regularize the angle parameters in the refinement network. To this end, we conduct extensive evaluations to validate that the proposed framework is powerful as it achieves state-of-the-art performance.

****

Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization With Spatially-Varying Lighting

Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, Matthias Niessner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3114-3122

We introduce a novel method to obtain high-quality 3D reconstructions from consumer RGB-D sensors. Our core idea is to simultaneously optimize for geometry encoded in a signed distance field (SDF), textures from automatically-selected keyframes, and their camera poses along with material and scene lighting. To this end, we propose a joint surface reconstruction approach that is based on Shape-from-Shading (SfS) techniques and utilizes the estimation of spatially-varying spherical harmonics (SVSH) from subvolumes of the reconstructed scene. Through extensive examples and evaluations, we demonstrate that our method dramatically increases the level of detail in the reconstructed scene geometry and contributes highly to consistent surface texture recovery.

****

Robust Hand Pose Estimation During the Interaction With an Unknown Object

Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, Karthik Ramani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3123-3132

This paper proposes a robust solution for accurate 3D hand pose estimation in the presence of an external object interacting with hands. Our main insight is that the shape of an object causes a configuration of the hand in the form of a hand grasp. Along this line, we simultaneously train deep neural networks using paired depth images. The object-oriented network learns functional grasps from an object perspective, whereas the hand-oriented network explores the details of hand configurations from a hand perspective. The two networks share intermediate observations produced from different perspectives to create a more informed representation. Our system then collaboratively classifies the grasp types and orientation of the hand and further constrains a pose space using these estimates. Finally, we collectively refine the unknown pose parameters to reconstruct the final hand pose. To this end, we conduct extensive evaluations to validate the efficacy of the proposed collaborative learning approach by comparing it with self-generated baselines and the state-of-the-art method.

****

Detailed Surface Geometry and Albedo Recovery From RGB-D Video Under Natural Illumination

Xinxin Zuo, Sen Wang, Jiangbin Zheng, Ruigang Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3133-3142

In this paper we present a novel approach for depth map enhancement from an RGB-D video sequence. The basic idea is to exploit the photometric information in the color sequence. Instead of making any assumption about surface albedo or controlled object motion and lighting, we use the lighting variations introduced by casual object movement. We are effectively calculating photometric stereo from a moving object under natural illuminations. The key technical challenge is to establish correspondences over the entire image set. We therefore develop a lighting insensitive robust pixel matching technique that out-performs optical flow method in presence of lighting variations. In addition we present an expectation-maximization framework to recover the surface normal and albedo simultaneously, without any regularization term. We have validated our method on both synthetic and real datasets to show its superior performance on both surface details recovery and intrinsic decomposition.

****

Monocular Free-Head 3D Gaze Tracking With Deep Learning and Geometry Constraints

Wangjiang Zhu, Haoping Deng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3143-3152

Free-head 3D gaze tracking outputs both the eye location and the gaze vector in 3D space, and it has wide applications in scenarios such as driver monitoring, advertisement analysis and surveillance. A reliable and low-cost monocular solution is critical for pervasive usage in these areas. Noticing that a gaze vector is a composition of head pose and eyeball movement in a geometrically deterministic way, we propose a novel gaze transform layer to connect separate head pose and eyeball movement models. The proposed decomposition does not suffer from head-gaze correlation overfitting and makes it possible to use datasets existing for other tasks. To add stronger supervision for better network training, we propose a two-step training strategy, which first trains sub-tasks with rough labels and then jointly trains with accurate gaze labels. To enable good cross-subject performance under various conditions, we collect a large dataset which has full coverage of head poses and eyeball movements, contains 200 subjects, and has diverse illumination conditions. Our deep solution achieves state-of-the-art gaze tracking accuracy, reaching 5.6 degrees cross-subject prediction error using a small network running at 1000 fps on a s ingle CPU (excluding face alignment time) and 4.3 degrees cross-subject error with a deeper network.

****

Filter Selection for Hyperspectral Estimation

Boaz Arad, Ohad Ben-Shahar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3153-3161

While recovery of hyperspectral signals from natural RGB images has been a recent subject of exploration, little to no consideration has been given to the camera response profiles used in the recovery process. In this paper we demonstrate that optimal selection of camera response filters may improve hyperspectral estimation accuracy by over 33%, emphasizing the importance of considering and selecting these response profiles wisely. Additionally, we present an evolutionary optimization methodology for optimal filter set selection from very large filter spaces, an approach that facilitates practical selection from families of customizable filters or filter optimization for multispectral cameras with more than 3 channels.

****

A Microfacet-Based Reflectance Model for Photometric Stereo With Highly Specular Surfaces

Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-Asa, Imari Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3162-3170

A precise, stable and invertible model for surface reflectance is the key to the success of photometric stereo with real world materials. Recent developments in the field have enabled shape recovery techniques for surfaces of various types, but an effective solution to directly estimating the surface normal in the pres ence of highly specular reflectance remains elusive. In this paper, we derive an analytical isotropic microfacet-based reflectance model, based on which a physically interpretable approximate is tailored for highly specular surfaces. With this approximate, we identify the equivalence between the surface recovery problem and the ellipsoid of revolution fitting problem, where the latter can be described as a system of polynomials. Additionally, we devise a fast, non-iterative and globally optimal solver for this problem. Experimental results on both synthetic and real images validate our model and demonstrate that our solution can stably deliver superior performance in its targeted application domain.

****

Detecting Faces Using Inside Cascaded Contextual CNN

Kaipeng Zhang, Zhanpeng Zhang, Hao Wang, Zhifeng Li, Yu Qiao, Wei Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3171-3179

Deep Convolutional Neural Networks (CNNs) achieve substantial improvements in face detection in the wild. Classical CNN-based face detection methods simply stack successive layers of filters where an input sample should pass through all layers before reaching a face/non-face decision. Inspired by the fact that for face detection, filters in deeper layers can discriminate between difficult face/non-face samples while those in shallower layers can efficiently reject simple non-face samples, we propose Inside Cascaded Structure that introduces face/non-face classifiers at different layers within the same CNN. In the training phase, we propose data routing mechanism which enables different layers to be trained by different types of samples, and thus deeper layers can focus on handling more difficult samples compared with traditional architecture. In addition, we introduce a two-stream contextual CNN architecture that leverages body part information adaptively to enhance face detection. Extensive experiments on the challenging FDDB and WIDER FACE benchmarks demonstrate that our method achieves competitive accuracy to the state-of-the-art techniques while keeps real time performance.

****

A Novel Space-Time Representation on the Positive Semidefinite Cone for Facial Expression Recognition

Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, Juan Carlos Alvarez-Paiva; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3180-3189

In this paper, we study the problem of facial expression recognition using a novel space-time geometric representation. We describe the temporal evolution of facial landmarks as parametrized trajectories on the Riemannian manifold of positive semidefinite matrices of fixed-rank. Our representation has the advantage to bring naturally a second desirable quantity when comparing shapes -- the spatial covariance -- in addition to the conventional affine-shape representation. We derive then geometric and computational tools for rate-invariant analysis and adaptive re-sampling of trajectories, grounding on the Riemannian geometry of the manifold. Specifically, our approach involves three steps: 1) facial landmarks are first mapped into the Riemannian manifold of positive semidefinite matrices of rank 2, to build time-parameterized trajectories; 2) a temporal alignment is performed on the trajectories, providing a geometry-aware (dis-)similarity measure between them; 3) finally, pairwise proximity function SVM (ppfSVM) is used to classify them, incorporating the latter (dis-)similarity measure into the kernel function. We show the effectiveness of the proposed approach on four publicly available benchmarks (CK+, MMI, Oulu-CASIA, and AFEW). The results of the proposed approach are comparable to or better than the state-of-the-art methods when involving only facial landmarks.

****

DeepCoder: Semi-Parametric Variational Autoencoders for Automatic Facial Action Coding

Dieu Linh Tran, Robert Walecki, Ognjen (Oggi) Rudovic, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3190-3199

Human face exhibits an inherent hierarchy in its representations (i.e., holistic facial expressions can be encoded via a set of facial action units (AUs) and their intensity). Variational (deep) auto-encoders (VAE) have shown great results in unsupervised extraction of hierarchical latent representations from large amounts of image data, while being robust to noise and other undesired artifacts. Potentially, this makes VAEs a suitable approach for learning facial features for AU intensity estimation. Yet, most existing VAE-based methods apply classifiers learned separately from the encoded features. By contrast, the non-parametric (probabilistic) approaches, such as Gaussian Processes (GPs), typically outperform their parametric counterparts, but cannot deal easily with large amounts of data. To this end, we propose a novel VAE semi-parametric modeling framework, named DeepCoder, which combines the modeling power of parametric (convolutional) and non-parametric (ordinal GPs) VAEs, for joint learning of (1) latent representations at multiple levels in a task hierarchy, and (2) classification of multiple ordinal outputs. We show on benchmark datasets for AU intensity estimation that the proposed DeepCoder outperforms the state-of-the-art approaches, and related VAEs and deep learning models.

****

Pose-Invariant Face Alignment With a Single CNN

Amin Jourabloo, Mao Ye, Xiaoming Liu, Liu Ren; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3200-3209

Face alignment has witnessed substantial progress in the last decade. One of the recent focuses has been aligning a dense 3D face shape to face images with large head poses. The dominant technology used is based on the cascade of regressors, e.g., CNNs, which has shown promising results. Nonetheless, the cascade of CNNs suffers from several drawbacks, e.g., lack of end-to-end training, hand-crafted features and slow training speed. To address these issues, we propose a new layer, named visualization layer, which can be integrated into the CNN architecture and enables joint optimization with different loss functions. Extensive evaluation of the proposed method on multiple datasets demonstrates state-of-the-art accuracy, while reducing the training time by more than half compared to the typical cascade of CNNs. In addition, we compare across multiple CNN architectures, all with the visualization layer, to further demonstrate the advantage of its utilization.

****

Unsupervised Domain Adaptation for Face Recognition in Unlabeled Videos

Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, Manmohan Chandraker; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3210-3218

Despite rapid advances in face recognition, there remains a clear gap between the performance of still image-based face recognition and video-based face recognition, due to the vast difference in visual quality between the domains and the difficulty of curating diverse large-scale video datasets. This paper addresses both of those challenges, through an image to video feature-level domain adaptation approach, to learn discriminative video frame representations. The framework utilizes large-scale unlabeled video data to reduce the gap between different domains while transferring discriminative knowledge from large-scale labeled still images. Given a face recognition network that is pretrained in the image domain, the adaptation is achieved by (i) distilling knowledge from the network to a video adaptation network through feature matching, (ii) performing feature restoration through synthetic data augmentation and (iii) learning a domain-invariant feature through a domain adversarial discriminator. We further improve performance through a discriminator-guided feature fusion that boosts high-quality frames while eliminating those degraded by video domain-specific factors. Experiments on the YouTube Faces and IJB-A datasets demonstrate that each module contributes to our feature-level domain adaptation framework and substantially improves video face recognition performance to achieve state-of-the-art accuracy. We demonstrate qualitatively that the network learns to suppress diverse artifacts in videos such as pose, illumination or occlusion without being explicitly trained for them.

****

Deeply-Learned Part-Aligned Representations for Person Re-Identification
Liming Zhao, Xi Li, Yueting Zhuang, Jingdong Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3219-3228
In this paper, we address the problem of person re-identification, which refers to associating the persons captured from different cameras. We propose a simple yet effective human part-aligned representation for handling the body part misalignment problem. Our approach decomposes the human body into regions (parts) which are discriminative for person matching, accordingly computes the representations over the regions, and aggregates the similarities computed between the corresponding regions of a pair of probe and gallery images as the overall matching score. Our formulation, inspired by attention models, is a deep neural network modeling the three steps together, which is learnt through minimizing the triplet loss function without requiring body part labeling information. Unlike most existing deep learning algorithms that learn a global or spatial partition-based local representation, our approach performs human body partition, and thus is more robust to pose changes and various human spatial distributions in the person bounding box. Our approach shows state-of-the-art results over standard datasets, Market-1501, CUHK03, CUHK01 and VIPeR.
****
Semantic Line Detection and Its Applications
Jun-Tae Lee, Han-Ul Kim, Chul Lee, Chang-Su Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3229-3237
Semantic lines characterize the layout of an image. Despite their importance in image analysis and scene understanding, there is no reliable research for semantic line detection. In this paper, we propose a semantic line detector using a convolutional neural network with multi-task learning, by regarding the line detection as a combination of classification and regression tasks. We use convolution and max-pooling layers to obtain multi-scale feature maps for an input image. Then, we develop the line pooling layer to extract a feature vector for each candidate line from the feature maps. Next, we feed the feature vector into the parallel classification and regression layers. The classification layer decides whether the line candidate is semantic or not. In case of a semantic line, the regression layer determines the offset for refining the line location. Experimental results show that the proposed detector extracts semantic lines accurately and reliably. Moreover, we demonstrate that the proposed detector can be used successfully in three applications: horizon estimation, composition enhancement, and image simplification.
****
A Generic Deep Architecture for Single Image Reflection Removal and Image Smoothing
Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, David Wipf; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3238-3247
This paper proposes a deep neural network structure that exploits edge information in addressing representative low-level vision tasks such as layer separation and image filtering. Unlike most other deep learning strategies applied in this context, our approach tackles these challenging problems by estimating edges and reconstructing images using only cascaded convolutional layers arranged such that no handcrafted or application-specific image-processing components are required. We apply the resulting transferrable pipeline to two different problem domains that are both sensitive to edges, namely, single image reflection removal and image smoothing. For the former, using a mild reflection smoothness assumption and a novel synthetic data generation method that acts as a type of weak supervision, our network is able to solve much more difficult reflection cases that cannot be handled by previous methods. For the latter, we also exceed the state-of-the-art quantitative and qualitative results by wide margins. In all cases, the proposed framework is simple, fast, and easy to transfer across disparate domains.
****
Revisiting Cross-Channel Information Transfer for Chromatic Aberration Correction

Tiancheng Sun, Yifan Peng, Wolfgang Heidrich; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3248-3256

Image aberrations can cause severe degradation in image quality for consumer-level cameras, especially under the current tendency to reduce the complexity of lens designs in order to shrink the overall size of modules. In simplified optical designs, chromatic aberration can be one of the most significant causes for degraded image quality, and it can be quite difficult to remove in post-processing, since it results in strong blurs in at least some of the color channels. In this work, we revisit the pixel-wise similarity between different color channels of the image and accordingly propose a novel algorithm for correcting chromatic aberration based on this cross-channel correlation. In contrast to recent weak prior-based models, ours uses strong pixel-wise fitting and transfer, which lead to significant quality improvements for large chromatic aberrations. Experimental results on both synthetic and real world images captured by different optical systems demonstrate that the chromatic aberration can be significantly reduced using our approach.

****

High-Quality Correspondence and Segmentation Estimation for Dual-Lens Smart-Phone Portraits

Xiaoyong Shen, Hongyun Gao, Xin Tao, Chao Zhou, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3257-3266

Estimating correspondence between two images and extracting the foreground object are two challenges in computer vision. With dual-lens smart phones, such as iPhone 7Plus and Huawei P9, coming into the market, two images of slightly different views provide us new information to unify the two topics. We propose a joint method to tackle them simultaneously via a joint fully connected conditional random field (CRF) framework. The regional correspondence is used to handle textureless regions in matching and make our CRF system computationally efficient. Our method is evaluated over 2,000 new image pairs, and produces promising results on challenging portrait images.

****

Learning Visual Attention to Identify People With Autism Spectrum Disorder

Ming Jiang, Qi Zhao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3267-3276

This paper presents a novel method for quantitative and objective diagnoses of Autism Spectrum Disorder (ASD) using eye tracking and deep neural networks. ASD is prevalent, with 1.5% of people in the US. The lack of clinical resources for early diagnoses has been a long-lasting issue. This work differentiates itself with three unique features: first, the proposed approach is data-driven and free of assumptions, important for new discoveries in understanding ASD as well as other neurodevelopmental disorders. Second, we concentrate our analyses on the differences in eye movement patterns between healthy people and those with ASD. An image selection method based on Fisher scores allows feature learning with the most discriminative contents, leading to efficient and accurate diagnoses. Third, we leverage the recent advances in deep neural networks for both prediction and visualization. Experimental results show the superior performance of our method in terms of multiple evaluation metrics used in diagnostic tests.

****

DSLR-Quality Photos on Mobile Devices With Deep Convolutional Networks

Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3277-3285

Despite a rapid rise in the quality of built-in smartphone cameras, their physical limitations - small sensor size, compact lenses and the lack of specific hardware, - impede them to achieve the quality results of DSLR cameras. In this work we present an end-to-end deep learning approach that bridges this gap by translating ordinary photos into DSLR-quality images. We propose learning the translation function using a residual convolutional neural network that improves both color rendition and image sharpness. Since the standard mean squared loss is not well suited for measuring perceptual image quality, we introduce a composite perc eptual error function that combines content, color and texture losses. The first two losses are defined analytically, while the texture loss is learned in an adversarial fashion. We also present DPED, a large-scale dataset that consists of real photos captured from three different phones and one high-end reflex camera. Our quantitative and qualitative assessments reveal that the enhanced image quality is comparable to that of DSLR-taken photos, while the methodology is generalized to any type of digital camera.

****

Non-Uniform Blind Deblurring by Reblurring

Yuval Bahat, Netalee Efrat, Michal Irani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3286-3294

We present an approach for blind image deblurring, which handles non-uniform blurs. Our algorithm has two main components: (i) A new method for recovering the unknown blur-field directly from the blurry image, and (ii) A method for deblurring the image given the recovered nonuniform blur-field. Our blur-field estimation is based on analyzing the spectral content of blurry image patches by Re-blurring them. Being unrestricted by any training data, it can handle a large variety of blur sizes, yielding superior blur-field estimation results compared to training based deep-learning methods. Our non-uniform deblurring algorithm is based on the internal image-specific patch recurrence prior. It attempts to recover a sharp image which, on one hand - results in the blurry image under our estimated blur-field, and on the other hand - maximizes the internal recurrence of patches within and across scales of the recovered sharp image. The combination of these two components gives rise to a blind-deblurring algorithm, which exceeds the performance of state-of-the-art CNN-based blind-deblurring by a significant margin, without the need for any training data.

****

Misalignment-Robust Joint Filter for Cross-Modal Image Pairs

Takashi Shibata, Masayuki Tanaka, Masatoshi Okutomi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3295-3304

Although several powerful joint filters for cross-modal image pairs have been proposed, the existing joint filters generate severe artifacts when there are misalignments between a target and a guidance images. Our goal is to generate an artifact-free output image even from the misaligned target and guidance images. We propose a novel misalignment-robust joint filter based on weight-volume-based image composition and joint-filter cost volume. Our proposed method first generates a set of translated guidances. Next, the joint-filter cost volume and a set of filtered images are computed from the target image and the set of the translated guidances. Then, a weight volume is obtained from the joint-filter cost volume while considering a spatial smoothness and a label-sparseness. The final output image is composed by fusing the set of the filtered images with the weight volume for the filtered images. The key is to generate the final output image directly from the set of the filtered images by weighted averaging using the weight volume that is obtained from the joint-filter cost volume. The proposed framework is widely applicable and can involve any kind of joint filter. Experimental results show that the proposed method is effective for various applications including image denosing, image up-sampling, haze removal and depth map interpolation.

****

Low-Rank Tensor Completion: A Pseudo-Bayesian Learning Approach

Wei Chen, Nan Song; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3305-3313

Low rank tensor completion, which solves a linear inverse problem with the principle of parsimony, is a powerful technique used in many application domains in computer vision and pattern recognition. As a surrogate function of the matrix rank that is non-convex and discontinuous, the nuclear norm is often used instead to derive efficient algorithms for recovering missing information in matrices and higher order tensors. However, the nuclear norm is a loose approximation of the matrix rank, and what is more, the tensor nuclear norm is not guaranteed to be the tightest convex envelope of a multilinear rank. Alternative algorithms either require specifying/tuning several parameters (e.g., the tensor rank), and/or

have a performance far from reaching the theoretical limit where the number of observed elements equals the degree of freedom in the unknown low-rank tensor. In this paper, we propose a pseudo-Bayesian approach, where a Bayesian-inspired cost function is adjusted using appropriate approximations that lead to desirable attributes including concavity and symmetry. Although deviating from the original Bayesian model, the resulting non-convex cost function is proved to have the ability to recover the true tensor with a low multilinear rank. A computational efficient algorithm is derived to solve the resulting non-convex optimization problem. We demonstrate the superior performance of the proposed algorithm in comparison with state-of-the-art alternatives by conducting extensive experiments on both synthetic data and several visual data recovery tasks.
****
DeepCD: Learning Deep Complementary Descriptors for Patch Representations
Tsun-Yi Yang, Jo-Han Hsu, Yen-Yu Lin, Yung-Yu Chuang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3314-3322
This paper presents the DeepCD framework which learns a pair of complementary descriptors jointly for a patch by employing deep learning techniques. It can be achieved by taking any descriptor learning architecture for learning a leading descriptor and augmenting the architecture with an additional network stream for learning a complementary descriptor. To enforce the complementary property, a new network layer, called data-dependent modulation (DDM) layer, is introduced for adaptively learning the augmented network stream with the emphasis on the training data that are not well handled by the leading stream. By optimizing the proposed joint loss function with late fusion, the obtained descriptors are complementary to each other and their fusion improves performance. Experiments on several problems and datasets show that the proposed method is simple yet effective, outperforming state-of-the-art methods.
****
Beyond Standard Benchmarks: Parameterizing Performance Evaluation in Visual Object Tracking
Luka Cehovin Zajc, Alan Lukezic, Ales Leonardis, Matej Kristan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3323-3331
Object-to-camera motion produces a variety of apparent motion patterns that significantly affect performance of short-term visual trackers. Despite being crucial for designing robust trackers, their influence is poorly explored in standard benchmarks due to weakly defined, biased and overlapping attribute annotations. In this paper we propose to go beyond pre-recorded benchmarks with post-hoc annotations by presenting an approach that utilizes omnidirectional videos to generate realistic, consistently annotated, short-term tracking scenarios with exactly parameterized motion patterns. We have created an evaluation system, constructed a fully annotated dataset of omnidirectional videos and generators for typical motion patterns. We provide an in-depth analysis of major tracking paradigms which is complementary to the standard benchmarks and confirms the expressiveness of our evaluation approach.
****
The Pose Knows: Video Forecasting by Generating Pose Futures
Jacob Walker, Kenneth Marino, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3332-3341
Current approaches to video forecasting attempt to generate videos directly in pixel space using Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs). However, since these approaches try to model all the structure and scene dynamics at once, in unconstrained settings they often generate uninterpretable results. Our insight is that forecasting needs to be done first at a higher level of abstraction. Specifically, we exploit human pose detectors as a free source of supervision and break the video forecasting problem into two discrete steps. First we explicitly model the high level structure of active objects in the scene (humans) and use a VAE to model the possible future movements of humans in the pose space. We then use the future poses generated as conditional information to a GAN to predict the future frames of the video in pixel space. By using the structured space of pose as an intermediate representation, we sidestep the

problems that GANs have in generating video pixels directly. We show through quantitative and qualitative evaluation that our method outperforms state-of-the-art methods for video prediction.
****

What Will Happen Next? Forecasting Player Moves in Sports Videos
Panna Felsen, Pulkit Agrawal, Jitendra Malik; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3342-3351
A large number of very popular team sports involve the act of one team trying to score a goal against the other. During this game play, defending players constantly try to predict the next move of the attackers to prevent them from scoring, whereas attackers constantly try to predict the next move of the defenders in order to defy them and score. Such behavior is a prime example of the general human faculty to make predictions about the future and is an important facet of human intelligence. An algorithmic solution to learning a model of the external world from sensory inputs in order to make forecasts is an important unsolved problem. In this work we develop a generic framework for forecasting future events in team sports videos directly from visual inputs. We introduce water polo and basketball datasets towards this end and compare the predictions of the proposed methods against expert and non-expert humans.
****

Robust Kronecker-Decomposable Component Analysis for Low-Rank Modeling
Mehdi Bahri, Yannis Panagakis, Stefanos Zafeiriou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3352-3361
Dictionary learning and component analysis are part of one of the most well-studied and active research fields, at the intersection of signal and image processing, computer vision, and statistical machine learning. In dictionary learning, the current methods of choice are arguably K-SVD and its variants, which learn a dictionary (i.e., a decomposition) for sparse coding via Singular Value Decomposition. In robust component analysis, leading methods derive from Principal Component Pursuit (PCP), which recovers a low-rank matrix from sparse corruptions of unknown magnitude and support. However, K-SVD is sensitive to the presence of noise and outliers in the training set. Additionally, PCP does not provide a dictionary that respects the structure of the data (e.g., images), and requires expensive SVD computations when solved by convex relaxation. In this paper, we introduce a new robust decomposition of images by combining ideas from sparse dictionary learning and PCP. We propose a novel Kronecker-decomposable component analysis which is robust to gross corruption, can be used for low-rank modeling, and leverages separability to solve significantly smaller problems. We design an efficient learning algorithm by drawing links with a restricted form of tensor factorization. The effectiveness of the proposed approach is demonstrated on real-world applications, namely background subtraction and image denoising, by performing a thorough comparison with the current state of the art.
****

Recurrent Topic-Transition GAN for Visual Paragraph Generation
Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, Eric P. Xing; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3362-3371
A natural image usually conveys rich semantic content and can be viewed from different angles. Existing image description methods are largely restricted by small sets of biased visual paragraph annotations, and fail to cover rich underlying semantics. In this paper, we investigate a semi-supervised paragraph generative framework that is able to synthesize diverse and semantically coherent paragraph descriptions by reasoning over local semantic regions and exploiting linguistic knowledge. The proposed Recurrent Topic-Transition Generative Adversarial Network (RTT-GAN) builds an adversarial framework between a structured paragraph generator and multi-level paragraph discriminators. The paragraph generator generates sentences recurrently by incorporating region-based visual and language attention mechanisms at each step. The quality of generated paragraph sentences is assessed by multi-level adversarial discriminators from two aspects, namely, plausibility at sentence level and topic-transition coherence at paragraph level. The joint adversarial training of RTT-GAN drives the model to generate realistic pa

ragraphs with smooth logical transition between sentence topics. Extensive quantitative experiments on image and video paragraph datasets demonstrate the effectiveness of our RTT-GAN in both supervised and semi-supervised settings. Qualitative results on telling diverse stories for an image verify the interpretability of RTT-GAN.
****

A Two-Streamed Network for Estimating Fine-Scaled Depth Maps From Single RGB Images

Jun Li, Reinhard Klein, Angela Yao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3372-3380

Estimating depth from a single RGB image is an ill-posed and inherently ambiguous problem. State-of-the-art deep learning methods can now estimate accurate 2D depth maps, but when the maps are projected into 3D, they lack local detail and are often highly distorted. We propose a fast-to-train two-streamed CNN that predicts depth and depth gradients, which are then fused together into an accurate and detailed depth map. We also define a novel set loss over multiple images; by regularizing the estimation between a common set of images, the network is less prone to over-fitting and achieves better accuracy than competing methods. Experiments on the NYU Depth v2 dataset shows that our depth predictions are competitive with state-of-the-art and lead to faithful 3D projections.
****

Weakly Supervised Object Localization Using Things and Stuff Transfer

Miaojing Shi, Holger Caesar, Vittorio Ferrari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3381-3390

We propose to help weakly supervised object localization for classes where location annotations are not available, by transferring things and stuff knowledge from a source set with available annotations. The source and target classes might share similar appearance (e.g. bear fur is similar to cat fur) or appear against similar background (e.g. horse and sheep appear against grass). To exploit this, we acquire three types of knowledge from the source set: a segmentation model trained on both thing and stuff classes; similarity relations between target and source classes; and co-occurrence relations between thing and stuff classes in the source. The segmentation model is used to generate thing and stuff segmentation maps on a target image, while the class similarity and co-occurrence knowledge help refining them. We then incorporate these maps as new cues into a multiple instance learning framework (MIL), propagating the transferred knowledge from the pixel level to the object proposal level. In extensive experiments, we conduct our transfer from the PASCAL Context dataset (source) to the ILSVRC, COCO and PASCAL VOC 2007 datasets (targets). We evaluate our transfer across widely different thing classes, including some that are not similar in appearance, but appear against similar background. The results demonstrate significant improvement over standard MIL, and we outperform the state-of-the-art in the transfer setting.
****

Single Image Action Recognition Using Semantic Body Part Actions

Zhichen Zhao, Huimin Ma, Shaodi You; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3391-3399

In this paper, we propose a novel single image action recognition algorithm based on the idea of semantic part actions. Unlike existing part-based methods, we argue that there exists a mid-level semantic, the semantic part action; and human action is a combination of semantic part actions and context cues. In detail, we divide human body into seven parts: head, torso, arms, hands and lower body. For each of them, we define a few semantic part actions (e.g.head: laughing). Finally, we exploit these part actions to infer the entire body action (e.g. applauding). To make the proposed idea practical, we propose a deep network-based framework which consists of two subnetworks, one for part localization and the other for action prediction. The action prediction network jointly learns part-level and body-level action semantics and combines them for the final decision. Extensive experiments demonstrate our proposal on semantic part actions as elements for entire body action. Our method reaches mAP of 93.9% and 91.2% on PASCAL VOC 20

12 and Stanford-40, which outperforms the state-of-the-art by 2.3% and 8.6%.

****

Incremental Learning of Object Detectors Without Catastrophic Forgetting
Konstantin Shmelkov, Cordelia Schmid, Karteek Alahari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3400-3409
Despite their success for object detection, convolutional neural networks are ill-equipped for incremental learning, i.e., adapting the original model trained on a set of classes to additionally detect objects of new classes, in the absence of the initial training data. They suffer from "catastrophic forgetting" - an abrupt degradation of performance on the original set of classes, when the training objective is adapted to the new classes. We present a method to address this issue, and learn object detectors incrementally, when neither the original training data nor annotations for the original classes in the new training set are available. The core of our proposed solution is a loss function to balance the interplay between predictions on the new classes and a new distillation loss which minimizes the discrepancy between responses for old classes from the original and the updated networks. This incremental learning can be performed multiple times, for a new set of classes in each step, with a moderate drop in performance compared to the baseline network trained on the ensemble of data. We present object detection results on the PASCAL VOC 2007 and COCO datasets, along with a detailed empirical analysis of the approach.

****

Generative Adversarial Networks Conditioned by Brain Signals
Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3410-3418
Recent advancements in generative adversarial networks (GANs), using deep convolutional models, have supported the development of image generation techniques able to reach satisfactory levels of realism. Further improvements have been proposed to condition GANs to generate images matching a specific object category or a short text description. In this work, we build on the latter class of approaches and investigate the possibility of driving and conditioning the image generation process by means of brain signals recorded, through an electroencephalograph (EEG), while users look at images from a set of 40 ImageNet object categories with the objective of generating the seen images. To accomplish this task, we first demonstrate that brain activity EEG signals encode visually-related information that allows us to accurately discriminate between visual object categories and, accordingly, we extract a more compact class-dependent representation of EEG data using recurrent neural networks. Afterwards, we use the learned EEG manifold to condition image generation employing GANs, which, during inference, will read EEG signals and convert them into images. We tested our generative approach using EEG signals recorded from six subjects while looking at images of the aforementioned 40 visual classes. The results show that for classes represented by well-defined visual patterns (e.g., pandas, airplane, etc.), the generated images are realistic and highly resemble those evoking the EEG signals used for conditioning GANs, resulting in an actual reading-the-mind process.

****

Learning to Disambiguate by Asking Discriminative Questions
Yining Li, Chen Huang, Xiaoou Tang, Chen Change Loy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3419-3428
The ability to ask questions is a powerful tool to gather information in order to learn about the world and resolve ambiguities. In this paper, we explore a novel problem of generating discriminative questions to help disambiguate visual instances. Our work can be seen as a complement and new extension to the rich research studies on image captioning and question answering. We introduce the first large-scale dataset with over 10,000 carefully annotated images-question tuples to facilitate benchmarking. In particular, each tuple consists of a pair of images and 4.6 discriminative questions (as positive samples) and 5.9 non-discriminative questions (as negative samples) on average. In addition, we present an effective method for visual discriminative question generation. The method can be tr

ained in a weakly supervised manner without discriminative images-question tuples but just existing visual question answering datasets. Promising results are shown against representative baselines through quantitative evaluations and user studies.

****

Interpretable Explanations of Black Boxes by Meaningful Perturbation

Ruth C. Fong, Andrea Vedaldi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3429-3437

As machine learning algorithms are increasingly applied to high impact yet high risk tasks, such as medical diagnosis or autonomous driving, it is critical that researchers can explain how such algorithms arrived at their predictions. In recent years, a number of image saliency methods have been developed to summarize where highly complex neural networks "look" in an image for evidence for their predictions. However, these techniques are limited by their heuristic nature and architectural constraints. In this paper, we make two main contributions: First, we propose a general framework for learning different kinds of explanations for any black box algorithm. Second, we specialise the framework to find the part of an image most responsible for a classifier decision. Unlike previous works, our method is model-agnostic and testable because it is grounded in explicit and interpretable image perturbations.

****

DeepRoadMapper: Extracting Road Topology From Aerial Images

Gellert Mattyus, Wenjie Luo, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3438-3446

Creating road maps is essential to the success of many applications such as autonomous driving and city planning. Most approaches in industry focus on leveraging expensive sensors mounted on top of a fleet of cars. This results in very accurate estimates when using techniques that involve a user in the loop. However, these solutions are very expensive and have small coverage. In contrast, in this paper we propose an approach that directly estimates road topology from aerial images. This provides us with an affordable solution which has large coverage. Towards this goal, we take advantage of the latest developments in deep learning to have an initial segmentation of the aerial images. We then propose an algorithm that reasons about missing connections in the extracted road topology as a shortest path problem which can be solved efficiently. We demonstrate the effectiveness of our approach in the challenging TorontoCity dataset and show very significant improvements over the state-of-the-art.

****

Monocular 3D Human Pose Estimation by Predicting Depth on Joints

Bruce Xiaohan Nie, Ping Wei, Song-Chun Zhu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3447-3455

This paper aims at estimating full-body 3D human poses from monocular images of which the biggest challenge is the inherent ambiguity introduced by lifting the 2D pose into 3D space. We propose a novel framework focusing on reducing this ambiguity by predicting the depth of human joints based on 2D human joint locations and body part images. Our approach is built on a two-level hierarchy of Long Short-Term Memory (LSTM) Networks which can be trained end-to-end. The first level consists of two components: 1) a skeleton-LSTM which learns the depth information from global human skeleton features; 2) a patch-LSTM which utilizes the local image evidence around joint locations. The both networks have tree structure defined on the kinematic relation of human skeleton, thus the information at different joints is broadcast through the whole skeleton in a top-down fashion. The two networks are first pre-trained separately on different data sources and then aggregated in the second layer for final depth prediction. The empirical evaluation on Human3.6M and HHOI dataset demonstrates the advantage of combining global 2D skeleton and local image patches for depth prediction, and our superior quantitative and qualitative performance relative to state-of-the-art methods.

****

Large-Scale Image Retrieval With Attentive Deep Local Features

Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, Bohyung Han; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3456-3465

We propose an attentive local feature descriptor suitable for large-scale image retrieval, referred to as DELF (DEep Local Feature). The new feature is based on convolutional neural networks, which are trained only with image-level annotations on a landmark image dataset. To identify semantically useful local features for image retrieval, we also propose an attention mechanism for keypoint selection, which shares most network layers with the descriptor. This framework can be used for image retrieval as a drop-in replacement for other keypoint detectors and descriptors, enabling more accurate feature matching and geometric verification. Our system produces reliable confidence scores to reject false positives---in particular, it is robust against queries that have no correct match in the database. To evaluate the proposed descriptor, we introduce a new large-scale dataset, referred to as Google-Landmarks dataset, which involves challenges in both database and query such as background clutter, partial occlusion, multiple landmarks, objects in variable scales, etc.

****

Deep Globally Constrained MRFs for Human Pose Estimation

Ioannis Marras, Petar Palasek, Ioannis Patras; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3466-3475

This work introduces a novel Convolutional Network architecture (ConvNet) for the task of human pose estimation, that is the localization of body joints in a single static image. We propose a coarse to fine architecture that addresses shortcomings of the baseline architecture in [26] that stem from the fact that large inaccuracies of its coarse ConvNet cannot be corrected by the refinement ConvNet that refines the estimation within small windows of the coarse prediction. We overcome this by introducing a Markov Random Field (MRF)-based spatial model network between the coarse and the refinement model that introduces geometric constraints on the relative locations of the body joints. We propose an architecture in which a) the filters that implement the message passing in the MRF inference are factored in a way that constrains them by a low dimensional pose manifold the projection to which is estimated by a separate branch of the proposed ConvNet and b) the strengths of the pairwise joint constraints are modeled by weights that are jointly estimated by the other parameters of the network. The proposed network is trained in an end-to-end fashion. Experimental results show that the proposed method improves the baseline model and provides state of the art results on very challenging benchmarks.

****

Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning

Soravit Changpinyo, Wei-Lun Chao, Fei Sha; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3476-3485

Leveraging class semantic descriptions and examples of known objects, zero-shot learning makes it possible to train a recognition model for an object class whose examples are not available. In this paper, we propose a novel zero-shot learning model that takes advantage of clustering structures in the semantic embedding space. The key idea is to impose the structural constraint that semantic representations must be predictive of the locations of their corresponding visual exemplars. To this end, this reduces to training multiple kernel-based regressors from semantic representation-exemplar pairs from labeled data of the seen object categories. Despite its simplicity, our approach significantly outperforms existing zero-shot learning methods in three out of four benchmark datasets, including the ImageNet dataset with more than 20,000 unseen categories.

****

Multi-Label Learning of Part Detectors for Heavily Occluded Pedestrian Detection

Chunluan Zhou, Junsong Yuan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3486-3495

Detecting pedestrians that are partially occluded remains a challenging problem due to variations and uncertainties of partial occlusion patterns. Following a commonly used framework of handling partial occlusions by part detection, we propose a multi-label learning approach to jointly learn part detectors to capture p artial occlusion patterns. The part detectors share a set of decision trees via boosting to exploit part correlations and also reduce the computational cost of applying these part detectors. The learned decision trees capture the overall distribution of all the parts. When used as a pedestrian detector individually, our part detectors learned jointly show better performance than their counterparts learned separately in different occlusion situations. The learned part detectors can be further integrated to better detect partially occluded pedestrians. Experiments on the Caltech dataset show state-of-the-art performance of our approach for detecting heavily occluded pedestrians.

****

SGN: Sequential Grouping Networks for Instance Segmentation

Shu Liu, Jiaya Jia, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3496-3504

In this paper, we propose Sequential Grouping Networks (SGN) to tackle the problem of object instance segmentation. SGNs employ a sequence of neural networks, each solving a sub-grouping problem of increasing semantic complexity in order to gradually compose objects out of pixels. In particular, the first network aims to group pixels along each image row and column by predicting horizontal and vertical object breakpoints. These breakpoints are then used to create line segments. By exploiting two-directional information, the second network groups horizontal and vertical lines into connected components. Finally, the third network groups the connected components into object instances. Our experiments show that our SGN significantly outperforms state-of-the-art approaches in both, the Cityscapes dataset as well as PASCAL VOC.

****

Adaptive Feeding: Achieving Fast and Accurate Detections by Adaptively Combining Object Detectors

Hong-Yu Zhou, Bin-Bin Gao, Jianxin Wu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3505-3513

Object detection aims at high speed and accuracy simultaneously. However, fast models are usually less accurate, while accurate models cannot satisfy our need for speed. A fast model can be 10 times faster but 50% less accurate than an accurate model. In this paper, we propose Adaptive Feeding (AF) to combine a fast (but less accurate) detector and an accurate (but slow) detector, by adaptively determining whether an image is easy or hard and choosing an appropriate detector for it. In practice, we build a cascade of detectors, including the AF classifier which make the easy vs. hard decision and the two detectors. The AF classifier can be tuned to obtain different tradeoff between speed and accuracy, which has negligible training time and requires no additional training data. Experimental results on the PASCAL VOC, MS COCO and Caltech Pedestrian datasets confirm that AF has the ability to achieve comparable speed as the fast detector and comparable accuracy as the accurate one at the same time. As an example, by combining the fast SSD300 with the accurate SSD500 detector, AF leads to 50% speedup over SSD500 with the same precision on the VOC2007 test set.

****

Aesthetic Critiques Generation for Photos

Kuang-Yu Chang, Kung-Hung Lu, Chu-Song Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3514-3523

It is said that a picture is worth a thousand words. Thus, there are various ways to describe an image, especially in aesthetic quality analysis. Although aesthetic quality assessment has generated a great deal of interest in the last decade, most studies focus on providing a quality rating of good or bad for an image. In this work, we extend the task to produce captions related to photo aesthetics and/or photography skills. To the best of our knowledge, this is the first study that deals with aesthetics captioning instead of AQ scoring. In contrast to common image captioning tasks that depict the objects or their relations in a picture, our approach can select a particular aesthetics aspect and generate captions with respect to the aspect chosen. Meanwhile, the proposed aspect-fusion method further uses an attention mechanism to generate more abundant aesthetics captions. We also introduce a new dataset for aesthetics captioning called the Photo

Critique Captioning Dataset (PCCD), which contains pair-wise image-comment data from professional photographers. The results of experiments on PCCD demonstrate that our approaches outperform existing methods for generating aesthetic-oriented captions for images.
****
Hide-And-Seek: Forcing a Network to Be Meticulous for Weakly-Supervised Object and Action Localization
Krishna Kumar Singh, Yong Jae Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3524-3533
We propose 'Hide-and-Seek', a weakly-supervised framework that aims to improve object localization in images and action localization in videos. Most existing weakly-supervised methods localize only the most discriminative parts of an object rather than all relevant parts, which leads to suboptimal performance. Our key idea is to hide patches in a training image randomly, forcing the network to seek other relevant parts when the most discriminative part is hidden. Our approach only needs to modify the input image and can work with any network designed for object localization. During testing, we do not need to hide any patches. Our Hide-and-Seek approach obtains superior performance compared to previous methods for weakly-supervised object localization on the ILSVRC dataset. We also demonstrate that our framework can be easily extended to weakly-supervised action localization.
****
Two-Phase Learning for Weakly Supervised Object Localization
Dahun Kim, Donghyeon Cho, Donggeun Yoo, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3534-3543
Weakly supervised semantic segmentation and localization have a problem of focusing only on the most important parts of an image since they use only image-level annotations. In this paper, we solve this problem fundamentally via two-phase learning. Our networks are trained in two steps. In the first step, a conventional fully convolutional network (FCN) is trained to find the most discriminative parts of an image. In the second step, the activations on the most salient parts are suppressed by inference conditional feedback, and then the second learning is performed to find the area of the next most important parts. By combining the activations of both phases, the entire portion of the target object can be captured. Our proposed training scheme is novel and can be utilized in well-designed techniques for weakly supervised semantic segmentation, salient region detection, and object location prediction. Detailed experiments demonstrate the effectiveness of our two-phase learning in each task.
****
Curriculum Dropout
Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, Rene Vidal, Vittorio Murino; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3544-3552
Dropout is a very effective way of regularizing neural networks. Stochastically dropping out units with a certain probability discourages over-specific co-adaptations of feature detectors, preventing overfitting and improving network generalization. Besides, Dropout can be interpreted as an approximate model aggregation technique, where an exponential number of smaller networks are averaged in order to get a more powerful ensemble. In this paper, we show that using a fixed dropout probability during training is a suboptimal choice. We thus propose a time scheduling for the probability of retaining neurons in the network. This induces an adaptive regularization scheme that smoothly increases the difficulty of the optimization problem. This idea of starting easy and adaptively increasing the difficulty of the learning problem has its roots in curriculum learning and allows one to train better models. Indeed, we prove that our optimization strategy implements a very general curriculum scheme, by gradually adding noise to both the input and intermediate feature representations within the network architecture. Experiments on seven image classification datasets and different network architectures show that our method, named Curriculum Dropout, frequently yields to better generalization and, at worst, performs just as well as the standard Dropou

t method.

****

Predictor Combination at Test Time

Kwang In Kim, James Tompkin, Christian Richardt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3553-3561

We present an algorithm for test-time combination of a set of reference predictors with unknown parametric forms. Existing multi-task and transfer learning algorithms focus on training-time transfer and combination, where the parametric forms of predictors are known and shared. However, when the parametric form of a predictor is unknown, e.g., for a human predictor or a predictor in a precompiled library, existing algorithms are not applicable. Instead, we empirically evaluate predictors on sampled data points to measure distances between different predictors. This embeds the set of reference predictors into a Riemannian manifold, upon which we perform manifold denoising to obtain the refined predictor. This allows our approach to make no assumptions about the underlying predictor forms. Our test-time combination algorithm equals or outperforms existing multi-task and transfer learning algorithms on challenging real-world datasets, without introducing specific model assumptions.

****

Guided Perturbations: Self-Corrective Behavior in Convolutional Neural Networks

Swami Sankaranarayanan, Arpit Jain, Ser Nam Lim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3562-3570

Convolutional Neural Networks have been a subject of great importance over the past decade and great strides have been made in their utility for producing state of the art performance in many computer vision problems. However, the behavior of deep networks is yet to be fully understood and is still an active area of research. In this work, we present an intriguing behavior: pre-trained CNNs can be made to improve their predictions by structurally perturbing the input. We observe that these perturbations - referred as Guided Perturbations - enable a trained network to improve its prediction performance without any learning or change in network weights. We perform various ablative experiments to understand how these perturbations affect the local context and feature representations. Furthermore, we demonstrate that this idea can improve performance of several existing approaches on semantic segmentation and scene labeling tasks on the PASCAL VOC dataset and supervised classification tasks on MNIST and CIFAR10 datasets.

****

Learning Robust Visual-Semantic Embeddings

Yao-Hung Hubert Tsai, Liang-Kang Huang, Ruslan Salakhutdinov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3571-3580

Many of the existing methods for learning joint embedding of images and text use only supervised information from paired images and its textual attributes. Taking advantage of the recent success of unsupervised learning in deep neural networks, we propose an end-to-end learning framework that is able to extract more robust multi-modal representations across domains. The proposed method combines representation learning models (i.e., auto-encoders) together with cross-domain learning criteria (i.e., Maximum Mean Discrepancy loss) to learn joint embeddings for semantic and visual features. A novel technique of unsupervised-data adaptation inference is introduced to construct more comprehensive embeddings for both labeled and unlabeled data. We evaluate our method on Animals with Attributes and Caltech-UCSD Birds 200-2011 dataset with a wide range of applications, including zero and few-shot image recognition and retrieval, from inductive to transductive settings. Empirically, we show that our framework improves over the current state of the art on many of the considered tasks.

****

PUnDA: Probabilistic Unsupervised Domain Adaptation for Knowledge Transfer Across Visual Categories

Behnam Gholami, Ognjen (Oggi) Rudovic, Vladimir Pavlovic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3581-3590

This paper introduces a probabilistic latent variable model to address unsupervised domain adaptation problems. This is achieved by learning projections from ea ch domain to a latent space along the classifier in the latent space to simultaneously minimizing a notion of domain disparity while maximizing a measure of discriminatory power. The non-parametric nature of our Latent variable model makes it possible to infer the latent space dimension automatically from data. We also develop a Variational Bayes (VB) algorithm for parameter estimation. We evaluate and contrast our proposed model against state-of-the-art methods for the task of visual domain adaptation using both handcrafted and deep net features. Our experiments show that even with a simple softmax classifier, our model can outperform several state-of-the-art methods taking advantage of more sophisticated classification schemes.

****

Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, Gregory D. Hager; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3591-3600

Many prediction tasks contain uncertainty. In some cases, uncertainty is inherent in the task itself. In future prediction, for example, many distinct outcomes are equally valid. In other cases, uncertainty arises from the way data is labeled. For example, in object detection, many objects of interest often go unlabeled, and in human pose estimation, occluded joints are often labeled with ambiguous values. In this work we focus on a principled approach for handling such scenarios. In particular, we propose a framework for reformulating existing single-prediction models as multiple hypothesis prediction (MHP) models and an associated meta loss and optimization procedure to train them. To demonstrate our approach, we consider four diverse applications: human pose estimation, future prediction, image classification and segmentation. We find that MHP models outperform their single-hypothesis counterparts in all cases, and that MHP models simultaneously expose valuable insights into the variability of predictions.

****

CDTS: Collaborative Detection, Tracking, and Segmentation for Online Multiple Object Segmentation in Videos

Yeong Jun Koh, Chang-Su Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3601-3609

A novel online algorithm to segment multiple objects in a video sequence is proposed in this work. We develop the collaborative detection, tracking, and segmentation (CDTS) technique to extract multiple segment tracks accurately. First, we jointly use object detector and tracker to generate multiple bounding box tracks for objects. Second, we transform each bounding box into a pixel-wise segment, by employing the alternate shrinking and expansion(ASE) segmentation. Third, we refine the segment tracks, by detecting object disappearance and reappearance cases and merging overlapping segment tracks. Experimental results show that the proposed algorithm significantly surpasses the state-of-the-art conventional algorithms on benchmark datasets.

****

Temporal Superpixels Based on Proximity-Weighted Patch Matching

Se-Ho Lee, Won-Dong Jang, Chang-Su Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3610-3618

A temporal superpixel algorithm based on proximity-weighted patch matching (TS-PPM) is proposed in this work. We develop the proximity-weighted patch matching (PPM), which estimates the motion vector of a superpixel robustly, by considering the patch matching distances of neighboring superpixels as well as the target superpixel. In each frame, we initialize superpixels by transferring the superpixel labels of the previous frame using PPM motion vectors. Then, we update the superpixel labels of boundary pixels, based on a cost function, composed of color, spatial, contour, and temporal consistency terms. Finally, we execute superpixel splitting, merging, and relabeling to regularize superpixel sizes and reduce incorrect labels. Experiments show that the proposed algorithm outperforms the state-of-the-art conventional algorithms significantly.

****

Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge

Ryota Hinami, Tao Mei, Shin'ichi Satoh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3619-3627

This paper addresses the problem of joint detection and recounting of abnormal events in videos. Recounting of abnormal events, i.e., explaining why they are judged to be abnormal, is an unexplored but critical task in video surveillance, because it helps human observers quickly judge if they are false alarms or not. To describe the events in the human-understandable form for event recounting, learning generic knowledge about visual concepts (e.g., object and action) is crucial. Although convolutional neural networks (CNNs) have achieved promising results in learning such concepts, it remains an open question as to how to effectively use CNNs for abnormal event detection, mainly due to the environment-dependent nature of the anomaly detection. In this paper, we tackle this problem by integrating a generic CNN model and environment-dependent anomaly detectors. Our approach first learns CNN with multiple visual tasks to exploit semantic information that is useful for detecting and recounting abnormal events. By appropriately plugging the model into anomaly detectors, we can detect and recount abnormal events while taking advantage of the discriminative power of CNNs. Our approach outperforms the state-of-the-art on Avenue and UCSD Ped2 benchmarks for abnormal event detection and also produces promising results of abnormal event recounting.

****

TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals

Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, Ram Nevatia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3628-3636

We address the problem of Temporal Action Proposal (TAP) generation. This is an important problem, as fast extraction of semantically important (e.g. human actions) segments from untrimmed videos is an important step for large-scale video analysis. To tackle this problem, we propose a novel Temporal Unit Regression Network (TURN) model. There are two salient aspects of TURN: (1) TURN jointly predicts action proposals and refines the temporal boundaries by temporal coordinate regression with contextual information; (2) Fast computation is enabled by unit feature reuse: a long untrimmed video is decomposed into video units, which are reused as basic building blocks of temporal proposals. TURN outperforms the state-of-the-art methods under average recall (AR) by a large margin on THUMOS-14 and ActivityNet datasets, and runs over 900 frames per second (FPS) on a TITAN X GPU. We further apply TURN as a proposal generation stage for existing temporal action localization pipelines, and outperforms state-of-the-art performance on THUMOS-14 and ActivityNet.

****

Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction

Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, Fabio Cuzzolin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3637-3646

We present a deep-learning framework for real-time multiple spatio-temporal (S/T) action localisation and classification. Current state-of-the-art approaches work offline, and are too slow to be useful in real-world settings. To overcome their limitations we introduce two major developments. Firstly, we adopt real-time SSD (Single Shot MultiBox Detector) CNNs to regress and classify detection boxes in each video frame potentially containing an action of interest. Secondly, we design an original and efficient online algorithm to incrementally construct and label "action tubes" from the SSD frame level detections. As a result, our system is not only capable of performing S/T detection in real time, but can also perform early action prediction in an online fashion. We achieve new state-of-the-art results in both S/T action localisation and early action prediction on the challenging UCF101-24 and J-HMDB-21 benchmarks, even when compared to the top offline competitors. To the best of our knowledge, ours is the first real-time (up to 40fps) system able to perform online S/T action localisation on the untrimmed videos of UCF101-24.

****

Leveraging Weak Semantic Relevance for Complex Video Event Classification

Chao Li, Jiewei Cao, Zi Huang, Lei Zhu, Heng Tao Shen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3647-3656

Existing video event classification approaches suffer from limited human-labeled semantic annotations. Weak semantic annotations can be harvested from Web-knowledge without involving any human interaction. However such weak annotations are noisy, thus can not be effectively utilized without distinguishing its reliability. In this paper, we propose a novel approach to automatically maximize the utility of weak semantic annotations (formalized as the semantic relevance of video shots to the target event) to facilitate video event classification. A novel attention model is designed to determine the attention scores of video shots, where the weak semantic relevance is considered as attentional guidance. Specifically, our model jointly optimizes two objectives at different levels. The first one is the classification loss corresponding to video-level groundtruth labels, and the second is the shot-level relevance loss corresponding to weak semantic relevance. We use a long short-term memory (LSTM) layer to capture the temporal information carried by the shots of a video. In each timestep, the LSTM employs the attention model to weight the current shot under the guidance of its weak semantic relevance to the event of interest. Thus, we can automatically exploit weak semantic relevance to assist video event classification. Extensive experiments have been conducted on three complex large-scale video event datasets i.e., MEDTest14, ActivityNet and FCVID. Our approach achieves the state-of-the-art classification performance on all three datasets. The significant performance improvement upon the conventional attention model also demonstrates the effectiveness of our model.

****

Weakly Supervised Summarization of Web Videos

Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, Amit K. Roy-Chowdhury; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3657-3666

Most of the prior works summarize videos by either exploring different heuristically designed criteria in an unsupervised way or developing fully supervised algorithms by leveraging human-crafted training data in form of video-summary pairs or importance annotations. However, unsupervised methods are blind to the video category and often fail to produce semantically meaningful video summaries. On the other hand, acquisition of large amount of training data in supervised approaches is non-trivial and may lead to a biased model. Different from existing methods, we introduce a weakly supervised approach that requires only video-level annotation for summarizing web videos. Casting the problem as a weakly supervised learning problem, we propose a flexible deep 3D CNN architecture to learn the notion of importance using only video-level annotation, and without any human-crafted training data. Specifically, our main idea is to leverage multiple videos of a category to automatically learn a parametric model for categorizing videos and then adopt the model to find important segments from a given video as the ones which have maximum influence to the model output. Furthermore, to unleash the full potential of our 3D CNN architecture, we also explored a series of good practices to reduce the influence of limited training data while summarizing videos. Experiments on two challenging and diverse datasets well demonstrate that our approach produces superior quality video summaries compared to several recently proposed approaches.

****

FCN-rLSTM: Deep Spatio-Temporal Neural Networks for Vehicle Counting in City Cameras

Shanghang Zhang, Guanhang Wu, Joao P. Costeira, Jose M. F. Moura; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3667-3676

In this paper, we develop deep spatio-temporal neural networks to sequentially count vehicles from low quality videos captured by city cameras (citycams). Citycam videos have low resolution, low frame rate, high occlusion and large perspective, making most existing methods lose their efficacy. To overcome limitations o f existing methods and incorporate the temporal information of traffic video, we design a novel FCN-rLSTM network to jointly estimate vehicle density and vehicle count by connecting fully convolutional neural networks (FCN) with long short term memory networks (LSTM) in a residual learning fashion. Such design leverages the strengths of FCN for pixel-level prediction and the strengths of LSTM for learning complex temporal dynamics. The residual learning connection reformulates the vehicle count regression as learning residual functions with reference to the sum of densities in each frame, which significantly accelerates the training of networks. To preserve feature map resolution, we propose a Hyper-Atrous combination to integrate atrous convolution in FCN and combine feature maps of different convolution layers. FCN-rLSTM enables refined feature representation and a novel end-to-end trainable mapping from pixels to vehicle count. We extensively evaluated the proposed method on different counting tasks with three datasets, with experimental results demonstrating their effectiveness and robustness. In particular, FCN-rLSTM reduces the mean absolute error (MAE) from 5.31 to 4.21 on TRANCOS; and reduces the MAE from 2.74 to 1.53 on WebCamT. Training process is accelerated by 5 times on average.

****

Fast Face-Swap Using Convolutional Neural Networks
Iryna Korshunova, Wenzhe Shi, Joni Dambre, Lucas Theis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3677-3685

We consider the problem of face swapping in images, where an input identity is transformed into a target identity while preserving pose, facial expression and lighting. To perform this mapping, we use convolutional neural networks trained to capture the appearance of the target identity from an unstructured collection of his/her photographs. This approach is enabled by framing the face swapping problem in terms of style transfer, where the goal is to render an image in the style of another one. Building on recent advances in this area, we devise a new loss function that enables the network to produce highly photorealistic results. By combining neural networks with simple pre- and post-processing steps, we aim at making face swap work in real-time with no input from the user.

****

Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images
Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3686-3695

With an increasing number of users sharing information online, privacy implications entailing such actions are a major concern. For explicit content, such as user profile or GPS data, devices (e.g. mobile phones) as well as web services (e.g. facebook) offer to set privacy settings in order to enforce the users' privacy preferences. We propose the first approach that extends this concept to image content in the spirit of a Visual Privacy Advisor. First, we categorize personal information in images into 68 image attributes and collect a dataset, which allows us to train models that predict such information directly from images. Second, we run a user study to understand the privacy preferences of different users w.r.t. such attributes. Third, we propose models that predict user specific privacy score from images in order to enforce the users' privacy preferences. Our model is trained to predict the user specific privacy risk and even outperforms the judgment of the users, who often fail to follow their own privacy preferences on image data.

****

First-Person Activity Forecasting With Online Inverse Reinforcement Learning
Nicholas Rhinehart, Kris M. Kitani; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3696-3705

We address the problem of incrementally modeling and forecasting long-term goals of a first-person camera wearer: what the user will do, where they will go, and what goal they seek. In contrast to prior work in trajectory forecasting, our algorithm, Darko, goes further to reason about semantic states (will I pick up an object?), and future goal states that are far both in terms of space and time. Darko learns and forecasts from first-person visual observations of the user's d aily behaviors via an Online Inverse Reinforcement Learning (IRL) approach. Classical IRL discovers only the rewards in a batch setting, whereas Darko discovers the states, transitions, rewards, and goals of a user from streaming data. Among other results, we show Darko forecasts goals better than competing methods in both noisy and ideal settings, and our approach is theoretically and empirically no-regret.

****

Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment With Limited Resources

Adrian Bulat, Georgios Tzimiropoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3706-3714

Our goal is to design architectures that retain the groundbreaking performance of CNNs for landmark localization and at the same time are lightweight, compact and suitable for applications with limited computational resources. To this end, we make the following contributions: (a) we are the first to study the effect of neural network binarization on localization tasks, namely human pose estimation and face alignment. We exhaustively evaluate various design choices, identify performance bottlenecks, and more importantly propose multiple orthogonal ways to boost performance. (b) Based on our analysis, we propose a novel hierarchical, parallel and multi-scale residual architecture that yields large performance improvement over the standard bottleneck block while having the same number of parameters, thus bridging the gap between the original network and its binarized counterpart. (c) We perform a large number of ablation studies that shed light on the properties and the performance of the proposed block. (d) We present results for experiments on the most challenging datasets for human pose estimation and face alignment, reporting in many cases state-of-the-art performance. Code can be downloaded from https://www.adrianbulat.com/binary-cnn-landmarks

****

MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, Christian Theobalt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3715-3724

In this work we propose a novel model-based deep convolutional autoencoder that addresses the highly challenging problem of reconstructing a 3D human face from a single in-the-wild color image. To this end, we combine a convolutional encoder network with an expert-designed generative model that serves as decoder. The core innovation is the differentiable parametric decoder that encapsulates image formation analytically based on a generative model. Our decoder takes as input a code vector with exactly defined semantic meaning that encodes detailed face pose, shape, expression, skin reflectance and scene illumination. Due to this new way of combining CNN-based with model-based face reconstruction, the CNN-based encoder learns to extract semantically meaningful parameters from a single monocular input image. For the first time, a CNN encoder and an expert-designed generative model can be trained end-to-end in an unsupervised manner, which renders training on very large (unlabeled) real world data feasible. The obtained reconstructions compare favorably to current state-of-the-art approaches in terms of quality and richness of representation.

****

RPAN: An End-To-End Recurrent Pose-Attention Network for Action Recognition in Videos

Wenbin Du, Yali Wang, Yu Qiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3725-3734

Recent studies demonstrate the effectiveness of Recurrent Neural Networks (RNNs) for action recognition in videos. However, previous works mainly utilize video-level category as supervision to train RNNs, which may prohibit RNNs to learn complex motion structures along time. In this paper, we propose a recurrent pose-attention network (RPAN) to address this challenge, where we introduce a novel pose-attention mechanism to adaptively learn pose-related features at every time-step action prediction of RNNs. More specifically, we make three main contributio ns in this paper. Firstly, unlike previous works on pose-related action recognition, our RPAN is an end-to-end recurrent network which can exploit important spatial-temporal evolutions of human pose to assist action recognition in a unified framework. Secondly, instead of learning individual human-joint features separately, our pose-attention mechanism learns robust human-part features by sharing attention parameters partially on the semantically-related human joints. These human-part features are then fed into the human-part pooling layer to construct a highly-discriminative pose-related representation for temporal action modeling. Thirdly, one important byproduct of our RPAN is pose estimation in videos, which can be used for coarse pose annotation in action videos. We evaluate the proposed RPAN quantitatively and qualitatively on two popular benchmarks, i.e., Sub-JHMDB and PennAction. Experimental results show that RPAN outperforms the recent state-of-the-art methods on these challenging datasets.

****

Temporal Non-Volume Preserving Approach to Facial Age-Progression and Age-Invariant Face Recognition

Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, Marios Savvides; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3735-3743

Modeling the long-term facial aging process is extremely challenging due to the presence of large and non-linear variations during the face development stages. In order to efficiently address the problem, this work first decomposes the aging process into multiple short-term stages. Then, a novel generative probabilistic model, named Temporal Non-Volume Preserving (TNVP) transformation, is presented to model the facial aging process at each stage. Unlike Generative Adversarial Networks (GANs), which requires an empirical balance threshold, and Restricted Boltzmann Machines (RBM), an intractable model, our proposed TNVP approach guarantees a tractable density function, exact inference and evaluation for embedding the feature transformations between faces in consecutive stages. Our model shows its advantages not only in capturing the non-linear age related variance in each stage but also producing a smooth synthesis in age progression across faces. Our approach can model any face in the wild provided with only four basic landmark points. Moreover, the structure can be transformed into a deep convolutional network while keeping the advantages of probabilistic models with tractable log-likelihood density estimation. Our method is evaluated in both terms of synthesizing age-progressed faces and cross-age face verification and consistently shows the state-of-the-art results in various face aging databases, i.e. FG-NET, MORPH, AginG Faces in the Wild (AGFW), and Cross-Age Celebrity Dataset (CACD). A large-scale face verification on Megaface challenge 1 is also performed to further show the advantages of our proposed approach.

****

Attribute-Enhanced Face Recognition With Neural Tensor Fusion Networks

Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S. Mukherjee, Timothy M. Hospedales, Neil M. Robertson, Yongxin Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3744-3753

Deep learning has achieved great success in face recognition, however deep-learned features still have limited invariance to strong intra-personal variations such as large pose. It is observed that some facial attributes (e.g. eyebrow thickness, gender) are invariant to such variations. We present the first work to systematically explore how the fusion of face recognition feature (FRF) and facial attribute feature (FAF) can enhance face recognition performance in various challenging scenarios. Despite this helpfulness of FAF, in practice, we find the existing fusion methods cannot reliably improve the recognition performance. Thus, we develop a powerful tensor-based framework which formulates this fusion as a low-rank tensor optimisation problem. It is non-trivial to directly optimise this tensor due to the large number of parameters to optimise. To solve this problem, we establish a theoretical equivalence between tensor optimisation and a two-stream gated neural network. This equivalence allows tractable computation and the use of standard neural network optimisation tools, leading to an accurate and stable optimisation. Experimental results show the fused feature works better th

an individual features thus proving for the first time that facial attributes aid face recognition. We achieve state-of-the-art performance on databases such as MultiPIE, CASIA NIR-VIR2.0 and LFW.
****

Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in Vitro

Zhedong Zheng, Liang Zheng, Yi Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3754-3762

The main contribution of this paper is a simple semi-supervised pipeline that only uses the original training set without collecting extra data. It is challenging in 1) how to obtain more training data only from the training set and 2) how to use the newly generated data. In this work, the generative adversarial network (GAN) is used to generate unlabeled samples. We propose the label smoothing regularization for outliers (LSRO). This method assigns a uniform label distribution to the unlabeled images, which regularizes the supervised model and improves the baseline. We verify the proposed method on a practical problem: person re-identification (re-ID). This task aims to retrieve a query person from other cameras. We adopt the deep convolutional generative adversarial network (DCGAN) for sample generation, and a baseline convolutional neural network (CNN) for representation learning. Experiments show that adding the GAN-generated data effectively improves the discriminative ability of learned CNN embeddings. On three large-scale datasets, Market-1501, CUHK03 and DukeMTMC-reID, we obtain +4.37%, +1.6% and +2.46% improvement in rank-1 precision over the baseline CNN, respectively. We additionally apply the proposed method to fine-grained bird recognition and achieve a +0.6% improvement over a strong baseline. The code is available at https://github.com/layumi/Person-reID_GAN.
****

Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks With Spatiotemporal Transformer Modules

Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, Jian Cheng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3763-3771

Gesture is a natural interface in interacting with wearable devices such as VR/AR helmet and glasses. The main challenge of gesture recognition in egocentric vision arises from the global camera motion caused by the spontaneous head movement of the device wearer. In this paper, we address the problem by a novel recurrent 3D convolutional neural network for end-to-end learning. We specially design a spatiotemporal transformer module with recurrent connections between neighboring time slices which can actively transform a 3D feature map into a canonical view in both spatial and temporal dimensions. To validate our method, we introduce a new dataset with sufficient size, variation and reality, which contains 83 gestures designed for interaction with wearable devices, and more than 24,000 RGB-D gesture samples from 50 subjects captured in 6 scenes. On this dataset, we show that the proposed network outperforms competing state-of-the-art algorithms. Moreover, our method can achieve state-of-the-art performance on the challenging GTEA egocentric action dataset.
****

Recursive Spatial Transformer (ReST) for Alignment-Free Face Recognition

Wanglong Wu, Meina Kan, Xin Liu, Yi Yang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3772-3780

Convolutional Neural Network (CNN) has led to significant progress in face recognition. Currently most CNN-based face recognition methods follow a two-step pipeline, i.e. a detected face is first aligned to a canonical one pre-defined by a mean face shape, and then it is fed into a CNN to extract features for recognition. The alignment step transforms all faces to the same shape, which can cause loss of geometrical information which is helpful in distinguishing different subjects. Moreover, it is hard to define a single optimal shape for the following recognition, since faces have large diversity in facial features, e.g. poses, illumination, etc. To be free from the above problems with an independent alignment step, we introduce a Recursive Spatial Transformer (ReST) module into CNN, allow ing face alignment to be jointly learned with face recognition in an end-to-end fashion. The designed ReST has an intrinsic recursive structure and is capable of progressively aligning faces to a canonical one, even those with large variations. To model non-rigid transformation, multiple ReST modules are organized in a hierarchical structure to account for different parts of faces. Overall, the proposed ReST can handle large face variations and non-rigid transformation, and is end-to-end learnable and adaptive to input, making it an effective alignment-free face recognition solution. Extensive experiments are performed on LFW and YTF datasets, and the proposed ReST outperforms those two-step methods, demonstrating its effectiveness.

****

Learning Discriminative Aggregation Network for Video-Based Face Recognition
Yongming Rao, Ji Lin, Jiwen Lu, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3781-3790

In this paper, we propose a discriminative aggregation network (DAN) for video face recognition, which aims to integrate information from video frames effectively and efficiently. Different from existing aggregation methods, our method aggregates raw video frames directly instead of the features obtained by complex processing. By combining the idea of metric learning and adversarial learning, we learn an aggregation network that produces more discriminative synthesized images compared to input frames. Our framework reduces the number of frames to be processed and greatly speed up the recognition procedure. Furthermore, low-quality frames containing misleading information are denoised during the aggregation process, making the system more robust and discriminative. Experimental results show that our framework can generate discriminative images from video clips and improve the overall recognition performance in both the speed and accuracy on three widely used datasets.

****

Synergy Between Face Alignment and Tracking via Discriminative Global Consensus Optimization
Muhammad Haris Khan, John McDonagh, Georgios Tzimiropoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3791-3799

An open question in facial landmark localization in video is whether one should perform tracking or tracking-by-detection (i.e. face alignment). Tracking produces fittings of high accuracy but is prone to drifting. Tracking-by-detection is drift-free but results in low accuracy fittings. To provide a solution to this problem, we describe the very first, to the best of our knowledge, synergistic approach between detection (face alignment) and tracking which completely eliminates drifting from face tracking, and does not merely perform tracking-by-detection. Our first main contribution is to show that one can achieve this synergy between detection and tracking using a principled optimization framework based on the theory of Global Variable Consensus Optimization using ADMM; Our second contribution is to show how the proposed analytic framework can be integrated within state-of-the-art discriminative methods for face alignment and tracking based on cascaded regression and deeply learned features. Overall, we call our method Discriminative Global Consensus Model (DGCM). Our third contribution is to show that DGCM achieves large performance improvement over the currently best performing face tracking methods on the most challenging category of the 300-VW dataset.

****

SVDNet for Pedestrian Retrieval
Yifan Sun, Liang Zheng, Weijian Deng, Shengjin Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3800-3808

This paper proposes the SVDNet for retrieval problems, with focus on the application of person re-identification (re-ID). We view each weight vector within a fully connected (FC) layer in a convolutional neuron network (CNN) as a projection basis. It is observed that the weight vectors are usually highly correlated. This problem leads to correlations among entries of the FC descriptor, and compromises the retrieval performance based on the Euclidean distance. To address the problem, this paper proposes to optimize the deep representation learning process with Singular Vector Decomposition (SVD). Specifically, with the restraint and relaxation iteration (RRI) training scheme, we are able to iteratively integrate the orthogonality constraint in CNN training, yielding the so-called SVDNet. We conduct experiments on the Market-1501, CUHK03, and Duke datasets, and show that RRI effectively reduces the correlation among the projection vectors, produces more discriminative FC descriptors, and significantly improves the re-ID accuracy. On the Market-1501 dataset, for instance, rank-1 accuracy is improved from 55.3% to 80.5% for CaffeNet, and from 73.8% to 82.3% for ResNet-50.

****

Towards More Accurate Iris Recognition Using Deeply Learned Spatially Corresponding Features

Zijing Zhao, Ajay Kumar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3809-3818

This paper proposes an accurate and generalizable deep learning framework for iris recognition. The proposed framework is based on a fully convolutional network (FCN), which generates spatially corresponding iris feature descriptors. A specially designed Extended Triplet Loss (ETL) function is introduced to incorporate the bit-shifting and non-iris masking, which are found necessary for learning discriminative spatial iris features. We also developed a sub-network to provide appropriate information for identifying meaningful iris regions, which serves as essential input for the newly developed ETL. Thorough experiments on four publicly available databases suggest that the proposed framework consistently outperforms several classic and state-of-the-art iris recognition approaches. More importantly, our model exhibits superior generalization capability as, unlike popular methods in the literature, it does not essentially require database-specific parameter tuning, which is another key advantage over other approaches.

****

Semantically Informed Multiview Surface Refinement

Maros Blaha, Mathias Rothermel, Martin R. Oswald, Torsten Sattler, Audrey Richard, Jan D. Wegner, Marc Pollefeys, Konrad Schindler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3819-3827

We present a method to jointly refine the geometry and semantic segmentation of 3D surface meshes. Our method alternates between updating the shape and the semantic labels. In the geometry refinement step, the mesh is deformed with variational energy minimization, such that it simultaneously maximizes photo-consistency and the compatibility of the semantic segmentations across a set of calibrated images. Label-specific shape priors account for interactions between the geometry and the semantic labels in 3D. In the semantic segmentation step, the labels on the mesh are updated with MRF inference, such that they are compatible with the semantic segmentations in the input images. Also, this step includes prior assumptions about the surface shape of different semantic classes. The priors induce a tight coupling, where semantic information influences the shape update and vice versa. Specifically, we introduce priors that favor (i) adaptive smoothing, depending on the class label; (ii) straightness of class boundaries; and (iii) semantic labels that are consistent with the surface orientation. The novel mesh-based reconstruction is evaluated in a series of experiments with real and synthetic data. We compare both to state-of-the-art, voxel-based semantic 3D reconstruction, and to purely geometric mesh refinement, and demonstrate that the proposed scheme yields improved 3D geometry as well as an improved semantic segmentation.

****

BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth

Mahdi Rad, Vincent Lepetit; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3828-3836

We introduce a novel method for 3D object detection and pose estimation from color images only. We first use segmentation to detect the objects of interest in 2D even in presence of partial occlusions and cluttered background. By contrast with recent patch-based methods, we rely on a "holistic" approach: We apply to the detected objects a Convolutional Neural Network (CNN) trained to predict their 3D poses in the form of 2D projections of the corners of their 3D bounding boxe

s. This, however, is not sufficient for handling objects from the recent T-LESS dataset: These objects exhibit an axis of rotational symmetry, and the similarity of two images of such an object under two different poses makes training the CNN challenging. We solve this problem by restricting the range of poses used for training, and by introducing a classifier to identify the range of a pose at run-time before estimating it. We also use an optional additional step that refines the predicted poses. We improve the state-of-the-art on the LINEMOD dataset from 73.7% to 89.3% of correctly registered RGB frames. We are also the first to report results on the Occlusion dataset using color images only. We obtain 54% of frames passing the Pose 6D criterion on average on several sequences of the T-LESS dataset, compared to the 67% of the state-of-the-art on the same sequences which uses both color and depth. The full approach is also scalable, as a single network can be trained for multiple objects simultaneously.

****

Modeling Urban Scenes From Pointclouds

William Nguatem, Helmut Mayer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3837-3846

We present a method for Modeling Urban Scenes from Pointclouds (MUSP). In contrast to existing approaches, MUSP is robust, scalable and provides a more complete description by not making a Manhattan-World assumption and modeling both buildings (with polyhedra) as well as the non-planar ground (using NURBS). First, we segment the scene into consistent patches using a divide-and-conquer based algorithm within a nonparametric Bayesian framework (stick-breaking construction). These patches often correspond to meaningful structures, such as the ground, facades, roofs and roof superstructures. We use polygon sweeping to fit predefined templates for buildings, and for the ground, a NURBS surface is fit and uniformly tessellated. Finally, we apply boolean operations to the polygons for buildings, buildings parts and the tesselated ground to clip unnecessary geometry (e.g., facades protrusions below the non-planar ground), leading to the final model. The explicit Bayesian formulation of scene segmentation makes our approach suitable for challenging datasets with varying amounts of noise, outliers, and point density. We demonstrate the robustness of MUSP on 3D pointclouds from image matching as well as LiDAR.

****

Parameter-Free Lens Distortion Calibration of Central Cameras

Filippo Bergamasco, Luca Cosmo, Andrea Gasparetto, Andrea Albarelli, Andrea Torsello; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3847-3855

At the core of many Computer Vision applications stands the need to define a mathematical model describing the imaging process. To this end, the pinhole model with radial distortion is probably the most commonly used, as it balances low complexity with a precision that is sufficient for most applications. On the other hand, unconstrained non-parametric models, despite being originally proposed to handle specialty cameras, have been shown to outperform the pinhole model, even with the simpler setups. Still, notwithstanding the higher accuracy, the inability of describing the imaging model by simple linear projective operators severely limits the use of standard algorithms with unconstrained models. In this paper we propose a parameter-free camera model where each imaging ray is constrained to a common optical center, forcing the camera to be central. Such model can be easily calibrated with a practical procedure which provides a convenient undistortion map that can be used to obtain a virtual pinhole camera. The proposed method can also be used to calibrate a stereo rig with a displacement map that simultaneously provides stereo rectification and corrects lens distortion.

****

Pose Guided RGBD Feature Learning for 3D Object Pose Estimation

Vassileios Balntas, Andreas Doumanoglou, Caner Sahin, Juil Sock, Rigas Kouskouridas, Tae-Kyun Kim; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3856-3864

In this paper we examine the effects of using object poses as guidance to learning robust features for 3D object pose estimation. Previous works have focused on learning feature embeddings based on metric learning with triplet comparisons and rely only on the qualitative distinction of similar and dissimilar pose labels. In contrast, we consider the exact pose differences between the training samples, and aim to learn embeddings such that the distances in the pose label space are proportional to the distances in the feature space. However, since it is less desirable to force the pose-feature correlation when objects are symmetric, we propose the data-driven weights that reflect object symmetry when measuring the pose distances. Furthermore, end-to-end pose regression is investigated and is shown to further boost the discriminative power of feature learning, improving pose recognition accuracies in NN, and thus can be used as another pose guidance to feature learning. Experimental results show that the features guided by poses, are significantly more discriminative than the ones learned in the traditional way, outperforming state-of-the-art works. Finally, we measure the generalisation capacities of pose guided feature learning in previously unseen scenes containing objects under different occlusion levels, and we show that it adapts well to novel tasks.

****

Efficient Global Illumination for Morphable Models

Andreas Schneider, Sandro Schonborn, Lavrenti Frobeen, Bernhard Egger, Thomas Vetter; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3865-3873

We propose an efficient self-shadowing illumination model for Morphable Models. Simulating self-shadowing with ray casting is computationally expensive which makes them impractical in Analysis-by-Synthesis methods for object reconstruction from single images. Therefore, we propose to learn self-shadowing for Morphable Model parameters directly with a linear model. Radiance transfer functions are a powerful way to represent self-shadowing used within the precomputed radiance transfer framework (PRT). We build on PRT to render deforming objects with self-shadowing at interactive frame rates. It can be illuminated efficiently by environment maps represented with spherical harmonics. The result is an efficient global illumination method for Morphable Models, exploiting an approximated radiance transfer. We apply the method to fitting Morphable Model parameters to a single image of a face and demonstrate that considering self-shadowing improves shape reconstruction.

****

Low Compute and Fully Parallel Computer Vision With HashMatch

Sean Ryan Fanello, Julien Valentin, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, Carlo Ciliberto, Philip Davidson, Shahram Izadi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3874-3883

Numerous computer vision problems such as stereo depth estimation, object-class segmentation and foreground/background segmentation can be formulated as per-pixel image labeling tasks. Given one or many images as input, the desired output of these methods is usually a spatially smooth assignment of labels. The large amount of such computer vision problems has lead to significant research efforts, with the state of art moving from CRF-based approaches to deep CNNs and more recently, hybrids of the two. Although these approaches have significantly advanced the state of the art, the vast majority has solely focused on improving quantitative results and are not designed for low-compute scenarios. In this paper, we present a new general framework for a variety of computer vision labeling tasks, called HashMatch. Our approach is designed to be both fully parallel, i.e. each pixel is independently processed, and low-compute, with a model complexity an order of magnitude less than existing CNN and CRF-based approaches. We evaluate HashMatch extensively on several problems such as disparity estimation, image retrieval, feature approximation and background subtraction, for which HashMatch achieves high computational efficiency while producing high quality results.

****

Dense Non-Rigid Structure-From-Motion and Shading With Unknown Albedos

Mathias Gallardo, Toby Collins, Adrien Bartoli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3884-3892

Significant progress has been recently made in Non-Rigid Structure-from-Motion (

NRSfM). However, existing methods do not handle poorly-textured surfaces that deform non-smoothly. These are nonetheless common occurrence in real-world applications. An important unanswered question is whether shading can be used to robustly handle these cases. Shading is complementary to motion because it constrains reconstruction densely at textureless regions, and has been used in several other reconstruction problems. The challenge we face is to simultaneously and densely estimate non-smooth, non-rigid shape from each image together with non-smooth, spatially-varying surface albedo (which is required to use shading). We tackle this using an energy-based formulation that combines a physical, discontinuity-preserving deformation prior with motion, shading and contour information. This is a largescale, highly non-convex optimization problem, and we propose a cascaded optimization that converges well without an initial estimate. Our approach works on both unorganized and organized small-sized image sets, and has been empirically validated on four real-world datasets for which all state-of-the-art approaches fail.

****

From Point Clouds to Mesh Using Regression

Lubor Ladicky, Olivier Saurer, SoHyeon Jeong, Fabio Maninchedda, Marc Pollefeys; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3893-3902

Surface reconstruction from a point cloud is a standard subproblem in many algorithms for dense 3D reconstruction from RGB images or depth maps. Methods, performing only local operations in the vicinity of individual points, are very fast, but reconstructed models typically contain lots of visually unpleasant holes. On the other hand, regularized volumetric approaches, formulated as a global optimization, are typically too slow for real-time interactive applications. We propose to use a regression forest based method, which predicts the projection of a grid point to the surface, depending on the spatial configuration of point density in the grid point neighborhood. We designed a suitable feature vector and efficient oct-tree based GPU evaluation, capable of predicting surface of high resolution 3D models in milliseconds. Our method learns and predicts surfaces from an observed point cloud sparser than the evaluation grid, and therefore effectively acts as a regularizer.

****

Stereo DSO: Large-Scale Direct Sparse Visual Odometry With Stereo Cameras

Rui Wang, Martin Schworer, Daniel Cremers; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3903-3911

We propose Stereo Direct Sparse Odometry (Stereo DSO) as a novel method for highly accurate real-time visual odometry estimation of large-scale environments from stereo cameras. It jointly optimizes for all the model parameters within the active window, including the intrinsic/extrinsic camera parameters of all keyframes and the depth values of all selected pixels. In particular, we propose a novel approach to integrate constraints from static stereo into the bundle adjustment pipeline of temporal multi-view stereo. Real-time optimization is realized by sampling pixels uniformly from image regions with sufficient intensity gradient. Fixed-baseline stereo resolves scale drift. It also reduces the sensitivities to large optical flow and to rolling shutter effect which are known shortcomings of direct image alignment methods. Quantitative evaluation demonstrates that the proposed Stereo DSO outperforms existing state-of-the-art visual odometry methods both in terms of tracking accuracy and robustness. Moreover, our method delivers a more precise metric 3D reconstruction than previous dense/semi-dense direct approaches while providing a higher reconstruction density than feature-based methods.

****

Space-Time Localization and Mapping

Minhaeng Lee, Charless C. Fowlkes; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3912-3921

This paper addresses the problem of building a spatio-temporal model of the world from a stream of time-stamped data. Unlike traditional models for simultaneous localization and mapping (SLAM) and structure-from-motion (SfM) which focus on

recovering a single rigid 3D model, we tackle the problem of mapping scenes in which dynamic components appear, move and disappear independently of each other over time. We introduce a simple generative probabilistic model of 4D structure which specifies location, spatial and temporal extent of rigid surface patches by local Gaussian mixtures. We fit this model to a time-stamped stream of input data using expectation-maximization to estimate the model structure parameters (mapping) and the alignment of the input data to the model (localization). By explicitly representing the temporal extent and observability of surfaces in a scene, our method yields superior localization and reconstruction relative to baselines that assume a static 3D scene. We carry out experiments on both synthetic RGB-D data streams as well as challenging real-world datasets, tracking scene dynamics in a human workspace over the course of several weeks.
****

Benchmarking Single-Image Reflection Removal Algorithms
Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Alex C. Kot; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3922-3930
Removing undesired reflections from a photo taken in front of a glass is of great importance for enhancing the efficiency of visual computing systems. Various approaches have been proposed and shown to be visually plausible on small datasets collected by their authors. A quantitative comparison of existing approaches using the same dataset has never been conducted due to the lack of suitable benchmark data with ground truth. This paper presents the first captured Single-image Reflection Removal dataset 'SIR2' with 40 controlled and 100 wild scenes, ground truth of background and reflection. For each controlled scene, we further provide ten sets of images under varying aperture settings and glass thicknesses. We perform quantitative and visual quality comparisons for four state-of-the-art singleimage reflection removal algorithms using four error metrics. Open problems for improving reflection removal algorithms are discussed at the end.
****

Attention-Aware Deep Reinforcement Learning for Video Face Recognition
Yongming Rao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3931-3940
In this paper, we propose an attention-aware deep reinforcement learning (ADRL) method for video face recognition, which aims to discard the misleading and confounding frames and find the focuses of attention in face videos for person recognition. We formulate the process of finding the attentions of videos as a Markov decision process and train the attention model through a deep reinforcement learning framework without using extra labels. Unlike existing attention models, our method takes information from both the image space and the feature space as the input to make better use of face information that is discarded in the feature learning process. Besides, our approach is attention-aware, which seeks different attentions of videos for the verification of different pairs of videos. Our approach achieves very competitive video face recognition performance on three widely used video face datasets.
****

Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation
Bugra Tekin, Pablo Marquez-Neila, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3941-3950
Most recent approaches to monocular 3D human pose estimation rely on Deep Learning. They typically involve regressing from an image to either 3D joint coordinates directly or 2D joint locations from which 3D coordinates are inferred. Both approaches have their strengths and weaknesses and we therefore propose a novel architecture designed to deliver the best of both worlds by performing both simultaneously and fusing the information along the way. At the heart of our framework is a trainable fusion scheme that learns how to fuse the information optimally instead of being hand-designed. This yields significant improvements upon the state-of-the-art on standard 3D human pose estimation benchmarks.
****

Deep Facial Action Unit Recognition From Partially Labeled Data
Shan Wu, Shangfei Wang, Bowen Pan, Qiang Ji; Proceedings of the IEEE Internation

al Conference on Computer Vision (ICCV), 2017, pp. 3951-3959

Current work on facial action unit (AU) recognition requires AU-labeled facial images. Although large amounts of facial images are readily available, AU annotation is expensive and time consuming. To address this, we propose a deep facial action unit recognition approach learning from partially AU-labeled data. The proposed approach makes full use of both partly available ground-truth AU labels and the readily available large scale facial images without annotation. Specifically, we propose to learn label distribution from the ground-truth AU labels, and then train the AU classifiers from the large-scale facial images by maximizing the log likelihood of the mapping functions of AUs with regard to the learnt label distribution for all training data and minimizing the error between predicted AUs and ground-truth AUs for labeled data simultaneously. A restricted Boltzmann machine is adopted to model AU label distribution, a deep neural network is used to learn facial representation from facial images, and the support vector machine is employed as the classifier. Experiments on two benchmark databases demonstrate the effectiveness of the proposed approach.

****

Pose-Driven Deep Convolutional Model for Person Re-Identification

Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3960-3969

Feature extraction and matching are two crucial components in person Re-Identification (ReID). The large pose deformations and the complex view variations exhibited by the captured person images significantly increase the difficulty of learning and matching of the features from person images. To overcome these difficulties, in this work we propose a Pose-driven Deep Convolutional (PDC) model to learn improved feature extraction and matching models from end to end. Our deep architecture explicitly leverages the human part cues to alleviate the pose variations and learn robust feature representations from both the global image and different local parts. To match the features from global human body and local body parts, a pose driven feature weighting sub-network is further designed to learn adaptive feature fusions. Extensive experimental analyses and results on three popular datasets demonstrate significant performance improvements of our model over all published stateof- the-art methods.

****

Recognition of Action Units in the Wild With Deep Nets and a New Global-Local Loss

C. Fabian Benitez-Quiroz, Yan Wang, Aleix M. Martinez; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3970-3979

Most previous algorithms for the recognition of Action Units (AUs) were trained on a small number of sample images. This was due to the limited amount of labeled data available at the time. This meant that data-hungry deep neural networks, which have shown their potential in other computer vision problems, could not be successfully trained to detect AUs. A recent publicly available database with close to a million labeled images has made this training possible. Image and individual variability (e.g., pose, scale, illumination, ethnicity) in this set is very large. Unfortunately, the labels in this dataset are not perfect (i.e., they are noisy), making convergence of deep nets difficult. To harness the richness of this dataset while being robust to the inaccuracies of the labels, we derive a novel global-local loss. This new loss function is shown to yield fast globally meaningful convergences and locally accurate results. Comparative results with those of the EmotioNet challenge demonstrate that our newly derived loss yields superior recognition of AUs than state-of-the-art algorithms.

****

Faster Than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses

Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, Marios Savvides; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3980-3989

Facial alignment involves finding a set of landmark points on an image with a kn

own semantic meaning. However, this semantic meaning of landmark points is often lost in 2D approaches where landmarks are either moved to visible boundaries or ignored as the pose of the face changes. In order to extract consistent alignment points across large poses, the 3D structure of the face must be considered in the alignment step. However, extracting a 3D structure from a single 2D image usually requires alignment in the first place. We present our novel approach to simultaneously extract the 3D shape of the face and the semantically consistent 2D alignment through a 3D Spatial Transformer Network (3DSTN) to model both the camera projection matrix and the warping parameters of a 3D model. By utilizing a generic 3D model and a Thin Plate Spline (TPS) warping function, we are able to generate subject specific 3D shapes without the need for a large 3D shape basis. In addition, our proposed network can be trained in an end-to-end framework on entirely synthetic data from the 300W-LP dataset. Unlike other 3D methods, our approach only requires one pass through the network resulting in a faster than real-time alignment. Evaluations of our model on the Annotated Facial Landmarks in the Wild (AFLW) and AFLW2000-3D datasets show our method achieves state-of-the-art performance over other 3D approaches to alignment.
****
Towards Large-Pose Face Frontalization in the Wild
Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, Manmohan Chandraker; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3990-3999
Despite recent advances in face recognition using deep learning, severe accuracy drops are observed for large pose variations in unconstrained environments. Learning pose-invariant features is one solution, but needs expensively labeled large-scale data and carefully designed feature learning algorithms. In this work, we focus on frontalizing faces in the wild under various head poses, including extreme profile views. We propose a novel deep 3D Morphable Model (3DMM) conditioned Face Frontalization Generative Adversarial Network (GAN), termed as FF-GAN, to generate neutral head pose face images. Our framework differs from both traditional GANs and 3DMM based modeling. Incorporating 3DMM into the GAN structure provides shape and appearance priors for fast convergence with less training data, while also supporting end-to-end training. The 3DMM-conditioned GAN employs not only the discriminator and generator loss but also a new masked symmetry loss to retain visual quality under occlusions, besides an identity loss to recover high frequency information. Experiments on face recognition, landmark localization and 3D reconstruction consistently show the advantage of our frontalization method on faces in the wild datasets.
****
A Joint Intrinsic-Extrinsic Prior Model for Retinex
Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, Dacheng Tao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4000-4009
We propose a joint intrinsic-extrinsic prior model to estimate both illumination and reflectance from an observed image. The 2D image formed from 3D object in the scene is affected by the intrinsic properties (shape and texture) and the extrinsic property (illumination). Based on a novel structure-preserving measure called local variation deviation, a joint intrinsic-extrinsic prior model is proposed for better prior representation. Better than conventional Retinex models, the proposed model can preserve the structure information by shape prior, estimate the reflectance with fine details by texture prior, and capture the luminous source by illumination prior. Experimental results demonstrate the effectiveness of the proposed method on simulated and real data. Compared with the other Retinex algorithms and state-of-the-art algorithms, the proposed model yields better results on both subjective and objective assessments.
****
Going Unconstrained With Rolling Shutter Deblurring
Mahesh Mohan M. R., A. N. Rajagopalan, Gunasekaran Seetharaman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4010-4018
Most present-day imaging devices are equipped with CMOS sensors. Motion blur is

a common artifact in hand-held cameras. Because CMOS sensors mostly employ a rolling shutter (RS), the motion deblurring problem takes on a new dimension. Although few works have recently addressed this problem, they suffer from many constraints including heavy computational cost, need for precise sensor information, and inability to deal with wide-angle systems (which most cell-phone and drone cameras are) and irregular camera trajectory. In this work, we propose a model for RS blind motion deblurring that mitigates these issues significantly. Comprehensive comparisons with state-of-the-art methods reveal that our approach not only exhibits significant computational gains and unconstrained functionality but also leads to improved deblurring performance.

****

A Stagewise Refinement Model for Detecting Salient Objects in Images

Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, Huchuan Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4019-4028

Deep convolutional neural networks (CNNs) have been successfully applied to a wide variety of problems in computer vision, including salient object detection. To detect and segment salient objects accurately, it is necessary to extract and combine high-level semantic features with low-level fine details simultaneously. This happens to be a challenge for CNNs as repeated subsampling operations such as pooling and convolution lead to a significant decrease in the initial image resolution, which results in loss of spatial details and finer structures. To remedy this problem, here we propose to augment feedforward neural networks with a novel pyramid pooling module and a multi-stage refinement mechanism for saliency detection. First, our deep feedward net is used to generate a coarse prediction map with much detailed structures lost. Then, refinement nets are integrated with local context information to refine the preceding saliency maps generated in the master branch in a stagewise manner. Further, a pyramid pooling module is applied for different region-based global context aggregation. Empirical evaluations over five benchmark datasets show that our proposed method compares favorably against the state-of-the-art approaches.

****

From Square Pieces to Brick Walls: The Next Challenge in Solving Jigsaw Puzzles

Shir Gur, Ohad Ben-Shahar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4029-4037

Research into computational jigsaw puzzle solving, an emerging theoretical problem with numerous applications, has focused in recent years on puzzles that constitute square pieces only. In this paper we wish to extend the scientific scope of appearance-based puzzle solving and consider "brick wall" jigsaw puzzles - rectangular pieces who may have different sizes, and could be placed next to each other at arbitrary offset along their abutting edge -- a more explicit configuration with propertie of real world puzzles. We present the new challenges that arise in brick wall puzzles and address them in two stages. First we concentrate on the reconstruction of the puzzle (with or without missing pieces) assuming an oracle for offset assignments. We show that despite the increased complexity of the problem, under these conditions performance can be made comparable to the state-of-the-art in solving the simpler square piece puzzles, and thereby argue that solving brick wall puzzles may be reduced to finding the correct offset between two neighboring pieces. We then move on to focus on implementing the oracle computationally using a mixture of dissimilarity metrics and correlation matching. We show results on various brick wall puzzles and discuss how our work may start a new research path for the puzzle solving community.

****

Online Video Deblurring via Dynamic Temporal Blending Network

Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, Michael Hirsch; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4038-4047

State-of-the-art video deblurring methods are capable of removing non-uniform blur caused by unwanted camera shake and/or object motion in dynamic scenes. However, most existing methods are based on batch processing and thus need access to all recorded frames, rendering them computationally demanding and time-consuming and thus limiting their practical use. In contrast, we propose an online (sequential) video deblurring method based on a spatio-temporal recurrent network that allows for real-time performance. In particular, we introduce a novel architecture which extends the receptive field while keeping the overall size of the network small to enable fast execution. In doing so, our network is able to remove even large blur caused by strong camera shake and/or fast moving objects. Furthermore, we propose a novel network layer that enforces temporal consistency between consecutive frames by dynamic temporal blending which compares and adaptively (at test time) shares features obtained at different time steps. We show the superiority of the proposed method in an extensive experimental evaluation.

****

Supervision by Fusion: Towards Unsupervised Learning of Deep Salient Object Detector

Dingwen Zhang, Junwei Han, Yu Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4048-4056

In light of the powerful learning capability of deep neural networks (DNNs), deep (convolutional) models have been built in recent years to address the task of salient object detection. Although training such deep saliency models can significantly improve the detection performance, it requires large-scale manual supervision in the form of pixel-level human annotation, which is highly labor-intensive and time-consuming. To address this problem, this paper makes the earliest effort to train a deep salient object detector without using any human annotation. The key insight is "supervision by fusion", i.e., generating useful supervisory signals from the fusion process of weak but fast unsupervised saliency models. Based on this insight, we combine an intra-image fusion stream and a inter-image fusion stream in the proposed framework to generate the learning curriculum and pseudo ground-truth for supervising the training of the deep salient object detector. Comprehensive experiments on four benchmark datasets demonstrate that our method can approach the same network trained with full supervision (within 2-5% performance gap) and, more encouragingly, even outperform a number of fully supervised state-of-the-art approaches.

****

Fast Multi-Image Matching via Density-Based Clustering

Roberto Tron, Xiaowei Zhou, Carlos Esteves, Kostas Daniilidis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4057-4066

We consider the problem of finding consistent matches across multiple images. Current state-of-the-art solutions use constraints on cycles of matches together with convex optimization, leading to computationally intensive iterative algorithms. In this paper, we instead propose a clustering-based formulation: we first rigorously show its equivalence with traditional approaches, and then propose QuickMatch, a novel algorithm that identifies multi-image matches from a density function in feature space. Specifically, QuickMatch uses the density estimate to order the points in a tree, and then extracts the matches by breaking this tree using feature distances and measures of distinctiveness. Our algorithm outperforms previous state-of-the-art methods (such as MatchALS) in accuracy, and it is significantly faster (up to 62 times faster on some benchmarks), and can scale to large datasets (with more than twenty thousands features).

****

Characterizing and Improving Stability in Neural Style Transfer

Agrim Gupta, Justin Johnson, Alexandre Alahi, Li Fei-Fei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4067-4076

Recent progress in style transfer on images has focused on improving the quality of stylized images and speed of methods. However, real-time methods are highly unstable resulting in visible flickering when applied to videos. In this work we characterize the instability of these methods by examining the solution set of the style transfer objective. We show that the trace of the Gram matrix representing style is inversely related to the stability of the method. Then, we present a recurrent convolutional network for real-time video style transfer which incorporates a temporal consistency loss and overcomes the instability of prior methods. Our networks can be applied at any resolution, do not require optical flow at test time, and produce high quality, temporally consistent stylized videos in real-time.

****

Cross-Modal Deep Variational Hashing

Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, Jie Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4077-4085

In this paper, we propose a cross-modal deep variational hashing (CMDVH) method to learn compact binary codes for cross-modality multimedia retrieval. Unlike most existing cross-modal hashing methods which learn a single pair of projections to map each example into a binary vector, we design a deep fusion neural network to learn non-linear transformations from image-text input pairs, such that a unified binary code is achieved in a discrete and discriminative manner using a classification-based hinge-loss criterion. We then design modality-specific neural networks in a probabilistic manner such that we model a latent variable to be close as possible from the inferred binary codes, at the same time approximated by a posterior distribution regularized by a known prior, which is suitable for out-of-sample extension. Experimental results on three benchmark datasets show the efficacy of the proposed approach.

****

Spatial Memory for Context Reasoning in Object Detection

Xinlei Chen, Abhinav Gupta; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4086-4096

Modeling instance-level context and object-object relationships is extremely challenging. It requires reasoning about bounding boxes of different locations, scales, aspect ratios etc.. Above all, instance-level spatial reasoning inherently requires modeling conditional distributions on previous detections. But our current object detection systems do not have any memory to remember what to condition on! The state-of-the-art object detectors still detect all object in parallel followed by non-maximal suppression (NMS). While memory has been used for tasks such as captioning and VQA, they use image-level memory cells without capturing the spatial layout. On the other hand, modeling object-object relationships requires spatial reasoning -- not only do we need a memory to store the spatial layout, but also a effective reasoning module to extract spatial patterns. This paper presents a conceptually simple yet powerful solution -- Spatial Memory Network (SMN), to model the instance-level context efficiently and effectively. Our spatial memory essentially assembles object instances back into a pseudo "image" representation that is easy to be fed into another ConvNet for object-object context reasoning. This leads to a new sequential reasoning architecture where image and memory are processed in parallel to obtain detections which update the memory again. We show our SMN architecture is effective as it provides 2.2% improvement over baseline Faster RCNN on the COCO dataset with VGG16.

****

Deep Binaries: Encoding Semantic-Rich Cues for Efficient Textual-Visual Cross Retrieval

Yuming Shen, Li Liu, Ling Shao, Jingkuan Song; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4097-4106

Cross-modal hashing is usually regarded as an effective technique for large-scale textual-visual cross retrieval, where data from different modalities are mapped into a shared Hamming space for matching. Most of the traditional textual-visual binary encoding methods only consider holistic image representations and fail to model descriptive sentences. This renders existing methods inappropriate to handle the rich semantics of informative cross-modal data for quality textual-visual search tasks. To address the problem of hashing cross-modal data with semantic-rich cues, in this paper, a novel integrated deep architecture is developed to effectively encode the detailed semantics of informative images and long descriptive sentences, named as Textual-Visual Deep Binaries (TVDB). In particular, region-based convolutional networks with long short-term memory units are introduced to fully explore image regional details while semantic cues of sentences are modeled by a text convolutional network. Additionally, we propose a stochastic batch-wise training routine, where high-quality binary codes and deep encoding functions are efficiently optimized in an alternating manner. Experiments are conducted on three multimedia datasets, i.e. Microsoft COCO, IAPR TC-12, and INRIA Web Queries, where the proposed TVDB model significantly outperforms state-of-the-art binary coding methods in the task of cross-modal retrieval.

****

Learning a Recurrent Residual Fusion Network for Multimodal Matching
Yu Liu, Yanming Guo, Erwin M. Bakker, Michael S. Lew; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4107-4116

A major challenge in matching between vision and language is that they typically have completely different features and representations. In this work, we introduce a novel bridge between the modality-specific representations by creating a co-embedding space based on a recurrent residual fusion (RRF) block. Specifically, RRF adapts the recurrent mechanism to residual learning, so that it can recursively improve feature embeddings while retaining the shared parameters. Then, a fusion module is used to integrate the intermediate recurrent outputs and generates a more powerful representation. In the matching network, RRF acts as a feature enhancement component to gather visual and textual representations into a more discriminative embedding space where it allows to narrow the cross-modal gap between vision and language. Moreover, we employ a bi-rank loss function to enforce separability of the two modalities in the embedding space. In the experiments, we evaluate the proposed RRF-Net using two multi-modal datasets where it achieves state-of-the-art results.

****

Rotational Subgroup Voting and Pose Clustering for Robust 3D Object Recognition
Anders Glent Buch, Lilita Kiforenko, Dirk Kraft; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4117-4125

It is possible to associate a highly constrained subset of relative 6 DoF poses between two 3D shapes, as long as the local surface orientation, the normal vector, is available at every surface point. Local shape features can be used to find putative point correspondences between the models due to their ability to handle noisy and incomplete data. However, this correspondence set is usually contaminated by outliers in practical scenarios, which has led to many past contributions based on robust detectors such as the Hough transform or RANSAC. The key insight of our work is that a single correspondence between oriented points on the two models is constrained to cast votes in a 1 DoF rotational subgroup of the full group of poses, SE(3). Kernel density estimation allows combining the set of votes efficiently to determine a full 6 DoF candidate pose between the models. This modal pose with the highest density is stable under challenging conditions, such as noise, clutter, and occlusions, and provides the output estimate of our method. We first analyze the robustness of our method in relation to noise and show that it handles high outlier rates much better than RANSAC for the task of 6 DoF pose estimation. We then apply our method to four state of the art data sets for 3D object recognition that contain occluded and cluttered scenes. Our method achieves perfect recall on two LIDAR data sets and outperforms competing methods on two RGB-D data sets, thus setting a new standard for general 3D object recognition using point cloud data.

****

CoupleNet: Coupling Global Structure With Local Parts for Object Detection
Yousong Zhu, Chaoyang Zhao, Jinqiao Wang, Xu Zhao, Yi Wu, Hanqing Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4126-4134

The region-based Convolutional Neural Network (CNN) detectors such as Faster R-CNN or R-FCN have already shown promising results for object detection by combining the region proposal subnetwork and the classification subnetwork together. Although R-FCN has achieved higher detection speed while keeping the detection performance, the global structure information is ignored by the position-sensitive score maps. To fully explore the local and global properties, in this paper, we propose a novel fully convolutional network, named as CoupleNet, to couple the global structure with local parts for object detection. Specifically, the object proposals obtained by the Region Proposal Network (RPN) are fed into the the cou pling module which consists of two branches. One branch adopts the position-sensitive RoI (PSRoI) pooling to capture the local part information of the object, while the other employs the RoI pooling to encode the global and context information. Next, we design different coupling strategies and normalization ways to make full use of the complementary advantages between the global and local branches. Extensive experiments demonstrate the effectiveness of our approach. We achieve state-of-the-art results on all three challenging datasets, i.e. a mAP of 82.7% on VOC07, 80.4% on VOC12, and 34.4% on COCO.Codes will be made publicly available.

****

Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, Bernt Schiele; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4135-4144

While strong progress has been made in image captioning recently, machine and human captions are still quite distinct. This is primarily due to the deficiencies in the generated word distribution, vocabulary size, and strong bias in the generators towards frequent captions. Furthermore, humans -- rightfully so -- generate multiple, diverse captions, due to the inherent ambiguity in the captioning task which is not explicitly considered in today's systems. To address these challenges, we change the training objective of the caption generator from reproducing ground-truth captions to generating a set of captions that is indistinguishable from human written captions. Instead of handcrafting such a learning target, we employ adversarial training in combination with an approximate Gumbel sampler to implicitly match the generated distribution to the human one. While our method achieves comparable performance to the state-of-the-art in terms of the correctness of the captions, we generate a set of diverse captions that are significantly less biased and better match the global uni-, bi- and tri-gram distributions of the human captions.

****

Drone-Based Object Counting by Spatially Regularized Regional Proposal Network

Meng-Ru Hsieh, Yen-Liang Lin, Winston H. Hsu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4145-4153

Existing counting methods often adopt regression-based approaches and cannot precisely localize the target objects, which hinders the further analysis (e.g., high-level understanding and fine-grained classification). In addition, most of prior work mainly focus on counting objects in static environments with fixed cameras. Motivated by the advent of unmanned flying vehicles (i.e., drones), we are interested in detecting and counting objects in such dynamic environments. We propose Layout Proposal Networks (LPNs) and spatial kernels to simultaneously count and localize target objects (e.g., cars) in videos recorded by the drone. Different from the conventional region proposal methods, we leverage the spatial layout information (e.g., cars often park regularly) and introduce these spatially regularized constraints into our network to improve the localization accuracy. To evaluate our counting method, we present a new large-scale car parking lot dataset (CARPK) that contains nearly 90,000 cars captured from different parking lots. To the best of our knowledge, it is the first and the largest drone view dataset that supports object counting, and provides the bounding box annotations.

****

BlitzNet: A Real-Time Deep Network for Scene Understanding

Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4154-4162

Real-time scene understanding has become crucial in many applications such as autonomous driving. In this paper, we propose a deep architecture, called BlitzNet, that jointly performs object detection and semantic segmentation in one forward pass, allowing real-time computations. Besides the computational gain of having a single network to perform several tasks, we show that object detection and semantic segmentation benefit from each other in terms of accuracy. Experimental results for VOC and COCO datasets show state-of-the-art performance for object detection and segmentation among real time systems.

****

Joint Learning of Object and Action Detectors

Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4163-4172

While most existing approaches for detection in videos focus on objects or human actions separately, we aim at jointly detecting objects performing actions, such as cat eating or dog jumping. We introduce an end-to-end multitask objective that jointly learns object-action relationships. We compare it with different training objectives, validate its effectiveness for detecting objects-actions in videos, and show that both tasks of object and action detection benefit from this joint learning. Moreover, the proposed architecture can be used for zero-shot learning of actions: our multitask objective leverages the commonalities of an action performed by different objects, eg. dog and cat jumping, enabling to detect actions of an object without training with these object-actions pairs. In experiments on the A2D dataset, we obtain state-of-the-art results on segmentation of object-action pairs. We finally apply our multitask architecture to detect visual relationships between objects in images of the VRD dataset.

****

Situation Recognition With Graph Neural Networks

Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4173-4182

We address the problem of recognizing situations in images. Given an image, the task is to predict the most salient verb (action), and fill its semantic roles such as who is performing the action, what is the source and target of the action, etc. Different verbs have different roles (e.g. attacking has weapon), and each role can take on many possible values (nouns). We propose a model based on Graph Neural Networks that allows us to efficiently capture joint dependencies between roles using neural networks defined on a graph. Experiments with different graph connectivities show that our approach that propagates information between roles significantly outperforms existing work, as well as multiple baselines. We obtain roughly 3-5% improvement over previous work in predicting the full situation. We also provide a thorough qualitative analysis of our model and influence of different roles in the verbs.

****

Learning Visual N-Grams From Web Data

Ang Li, Allan Jabri, Armand Joulin, Laurens van der Maaten; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4183-4192

Real-world image recognition systems need to recognize tens of thousands of classes that constitute a plethora of visual concepts. The traditional approach of annotating thousands of images per class for training is infeasible in such a scenario, prompting the use of webly supervised data. This paper explores the training of image-recognition systems on large numbers of images and associated user comments. In particular, we develop visual n-gram models that can predict arbitrary phrases that are relevant to the content of an image. Our visual n-gram models are feed-forward convolutional networks trained using new loss functions that are inspired by n-gram models commonly used in language modeling. We demonstrate the merits of our models in phrase prediction, phrase-based image retrieval, relating images and captions, and zero-shot transfer.

****

Attention-Based Multimodal Fusion for Video Description

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, Kazuhiko Sumi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4193-4202

Current methods for video description are based on encoder-decoder sentence generation using recurrent neural networks (RNNs). Recent work has demonstrated the advantages of integrating temporal attention mechanisms into these models, in wh ich the decoder network predicts each word in the description by selectively giving more weight to encoded features from specific time frames. Such methods typically use two different types of features: image features (from an object classification model), and motion features (from an action recognition model), combined by naive concatenation in the model input. Because different feature modalities may carry task-relevant information at different times, fusing them by naive concatenation may limit the model's ability to dynamically determine the relevance of each type of feature to different parts of the description. In this paper, we incorporate audio features in addition to the image and motion features. To fuse these three modalities, we introduce a multimodal attention model that can selectively utilize features from different modalities for each word in the output description. Combining our new multimodal attention model with standard temporal attention outperforms state-of-the-art methods on two standard datasets: YouTube2Text and MSR-VTT.

****

Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding From Fashion Images

Wei-Lin Hsiao, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4203-4212

What defines a visual style? Fashion styles emerge organically from how people assemble outfits of clothing, making them difficult to pin down with a computational model. Low-level visual similarity can be too specific to detect stylistically similar images, while manually crafted style categories can be too abstract to capture subtle style differences. We propose an unsupervised approach to learn a style-coherent representation. Our method leverages probabilistic polylingual topic models based on visual attributes to discover a set of latent style factors. Given a collection of unlabeled fashion images, our approach mines for the latent styles, then summarizes outfits by how they mix those styles. Our approach can organize galleries of outfits by style without requiring any style labels. Experiments on over 100K images demonstrate its promise for retrieving, mixing, and summarizing fashion images by their style.

****

Aligned Image-Word Representations Improve Inductive Transfer Across Vision-Language Tasks

Tanmay Gupta, Kevin Shih, Saurabh Singh, Derek Hoiem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4213-4222

An important goal of computer vision is to build systems that learn visual representations over time that can be applied to many tasks. In this paper, we investigate a vision-language embedding as a core representation and show that it leads to better cross-task transfer than standard multi-task learning. In particular, the task of visual recognition is aligned to the task of visual question answering by forcing each to use the same word-region embeddings. We show this leads to greater inductive transfer from recognition to VQA than standard multitask learning. Visual recognition also improves, especially for categories that have relatively few recognition training labels but appear often in the VQA setting. Thus, our paper takes a small step towards creating more general vision systems by showing the benefit of interpretable, flexible, and trainable core representations.

****

Learning Discriminative Latent Attributes for Zero-Shot Classification

Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, Xilin Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4223-4232

Zero-shot learning (ZSL) aims to transfer knowledge from observed classes to the unseen classes, based on the assumption that both the seen and unseen classes share a common semantic space, among which attributes enjoy a great popularity. However, few works study whether the human-designed semantic attributes are discriminative enough to recognize different classes. Moreover, attributes are often correlated with each other, which makes it less desirable to learn each attribute independently. In this paper, we propose to learn a latent attribute space, which is not only discriminative but also semantic-preserving, to perform the ZSL task. Specifically, a dictionary learning framework is exploited to connect the latent attribute space with attribute space and similarity space. Extensive experiments on four benchmark datasets show the effectiveness of the proposed approach.

****

PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN

Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, Shih-Fu Chang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4233-4241

We aim to tackle a novel vision task called Weakly Supervised Visual Relation Detection (WSVRD) to detect "subject-predicate-object" relations in an image with object relation groundtruths available only at the image level. This is motivated by the fact that it is extremely expensive to label the combinatorial relations between objects at the instance level. Compared to the extensively studied problem, Weakly Supervised Object Detection (WSOD), WSVRD is more challenging as it needs to examine a large set of regions pairs, which is computationally prohibitive and more likely stuck in a local optimal solution such as those involving wrong spatial context. To this end, we present a Parallel, Pairwise Region-based, Fully Convolutional Network (PPR-FCN) for WSVRD. It uses a parallel FCN architecture that simultaneously performs pair selection and classification of single regions and region pairs for object and relation detection, while sharing almost all computation shared over the entire image. In particular, we propose a novel position-role-sensitive score map with pairwise RoI pooling to efficiently capture the crucial context associated with a pair of objects. We demonstrate the superiority of PPR-FCN over all baselines in solving the WSVRD challenge by using results of extensive experiments over two visual relation benchmarks.

****

Higher-Order Minimum Cost Lifted Multicuts for Motion Segmentation

Margret Keuper; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4242-4250

Most state-of-the-art motion segmentation algorithms draw their potential from modeling motion differences of local entities such as point trajectories in terms of pairwise potentials in graphical models. Inference in instances of minimum cost multicut problems defined on such graphs allows to optimize the number of the resulting segments along with the segment assignment. However, pairwise potentials limit the discriminative power of the employed motion models to translational differences. More complex models such as Euclidean or affine transformations call for higher-order potentials and a tractable inference in the resulting higher-order graphical models. In this paper, we (1) introduce a generalization of the minimum cost lifted multicut problem to hypergraphs, and (2) propose a simple primal feasible heuristic that allows for a reasonably efficient inference in instances of higher-order lifted multicut problem instances defined on point trajectory hypergraphs for motion segmentation. The resulting motion segmentations improve over the state-of-the-art on the FBMS-59 dataset.

****

Deep Free-Form Deformation Network for Object-Mask Registration

Haoyang Zhang, Xuming He; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4251-4259

This paper addresses the problem of object-mask registration, which aligns a shape mask to a target object instance. Prior work typically formulate the problem as an object segmentation task with mask prior, which is challenging to solve. In this work, we take a transformation based approach that predicts a 2D non-rigid spatial transform and warps the shape mask onto the target object. In particular, we propose a deep spatial transformer network that learns free-form deformations (FFDs) to non-rigidly warp the shape mask based on a multi-level dual mask feature pooling strategy. The FFD transforms are based on B-splines and parameterized by the offsets of predefined control points, which are differentiable. Therefore, we are able to train the entire network in an end-to-end manner based on L2 matching loss. We evaluate our FFD network on a challenging object-mask alignment task, which aims to refine a set of object segment proposals, and our approach achieves the state-of-the-art performance on the Cityscapes, the PASCAL VOC and the MSCOCO datasets.

****

Region-Based Correspondence Between 3D Shapes via Spatially Smooth Biclustering

Matteo Denitto, Simone Melzi, Manuele Bicego, Umberto Castellani, Alessandro Farinelli, Mario A. T. Figueiredo, Yanir Kleiman, Maks Ovsjanikov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4260-4269

Region-based correspondence (RBC) is a highly relevant and non-trivial computer vision problem. Given two 3D shapes, RBC seeks segments/regions on these shapes that can be reliably put in correspondence. The problem thus consists both in finding the regions and determining the correspondences between them. This problem statement is similar to that of "biclustering", implying that RBC can be cast as a biclustering problem. Here, we exploit this implication by tackling RBC via a novel biclustering approach, called S4B (spatially smooth spike and slab biclustering), which: (i) casts the problem in a probabilistic low-rank matrix factorization perspective; (ii) uses a spike and slab prior to induce sparsity; (iii) is enriched with a spatial smoothness prior, based on geodesic distances, encouraging nearby vertices to belong to the same bicluster. This type of spatial prior cannot be used in classical biclustering techniques. We test the proposed approach on the FAUST dataset, outperforming both state-of-the-art RBC techniques and classical biclustering methods.

****

Learning Discriminative ab-Divergences for Positive Definite Matrices

Anoop Cherian, Panagiotis Stanitsas, Mehrtash Harandi, Vassilios Morellas, Nikolaos Papanikolopoulos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4270-4279

Symmetric positive definite (SPD) matrices are useful for capturing second-order statistics of visual data. To compare two SPD matrices, several measures are available, such as the affine-invariant Riemannian metric, Jeffreys divergence, Jensen-Bregman logdet divergence, etc.; however, their behaviors may be application dependent, raising the need of manual selection to achieve the best possible performance. Further and as a result of their overwhelming complexity for large-scale problems, computing pairwise similarities by clever embedding of SPD matrices is often preferred to direct use of the aforementioned measures. In this paper, we propose a discriminative metric learning framework, Information Divergence and Dictionary Learning (IDDL), that not only learns application specific measures on SPD matrices automatically, but also embeds them as vectors using a learned dictionary. To learn the similarity measures (which could potentially be distinct for every dictionary atom), we use the recently introduced alpha-beta-logdet divergence, which is known to unify the measures listed above. We propose a novel IDDL objective, that learns the parameters of the divergence and the dictionary atoms jointly in a discriminative setup and is solved efficiently using Riemannian optimization. We showcase extensive experiments on eight computer vision datasets, demonstrating state-of-the-art performances.

****

Consensus Convolutional Sparse Coding

Biswarup Choudhury, Robin Swanson, Felix Heide, Gordon Wetzstein, Wolfgang Heidrich; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4280-4288

Convolutional sparse coding (CSC) is a promising direction for unsupervised learning in computer vision. In contrast to recent supervised methods, CSC allows for convolutional image representations to be learned that are equally useful for high-level vision tasks and low-level image reconstruction and can be applied to a wide range of tasks without problem-specific retraining. Due to their extreme memory requirements, however, existing CSC solvers have so far been limited to low-dimensional problems and datasets using a handful of low-resolution example images at a time. In this paper, we propose a new approach to solving CSC as a consensus optimization problem, which lifts these limitations. By learning CSC features from large-scale image datasets for the first time, we achieve significant quality improvements in a number of imaging tasks. Moreover, the proposed method enables new applications in high-dimensional feature learning that has been i ntractable using existing CSC methods. This is demonstrated for a variety of reconstruction problems across diverse problem domains, including 3D multispectral demosaicing and 4D light field view synthesis.

****

Domain-Adaptive Deep Network Compression

Marc Masana, Joost van de Weijer, Luis Herranz, Andrew D. Bagdanov, Jose M. Alvarez; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4289-4297

Deep Neural Networks trained on large datasets can be easily transferred to new domains with far fewer labeled examples by a process called fine-tuning. This has the advantage that representations learned in the large source domain can be exploited on smaller target domains. However, networks designed to be optimal for the source task are often prohibitively large for the target task. In this work we address the compression of networks after domain transfer. We focus on compression algorithms based on low-rank matrix decomposition. Existing methods base compression solely on learned network weights and ignore the statistics of network activations. We show that domain transfer leads to large shifts in network activations and that it is desirable to take this into account when compressing. We demonstrate that considering activation statistics when compressing weights leads to a rank-constrained regression problem with a closed-form solution. Because our method takes into account the target domain, it can more optimally remove the redundancy in the weights. Experiments show that our Domain Adaptive Low Rank (DALR) method significantly outperforms existing low-rank compression techniques. With our approach, the fc6 layer of VGG19 can be compressed more than 4x more than using truncated SVD alone -- with only a minor or no loss in accuracy. When applied to domain-transferred networks it allows for compression down to only 5-20% of the original number of parameters with only a minor drop in performance.

****

Self-Supervised Learning of Pose Embeddings From Spatiotemporal Relations in Videos

Omer Sumer, Tobias Dencker, Bjorn Ommer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4298-4307

Human pose analysis is presently dominated by deep convolutional networks trained with extensive manual annotations of joint locations and beyond. To avoid the need for expensive labeling, we exploit spatiotemporal relations in training videos for self-supervised learning of pose embeddings. The key idea is to combine temporal ordering and spatial placement estimation as auxiliary tasks for learning pose similarities in a Siamese convolutional network. Since the self-supervised sampling of both tasks from natural videos can result in ambiguous and incorrect training labels, our method employs a curriculum learning idea that starts training with the most reliable data samples and gradually increases the difficulty. To further refine the training process we mine repetitive poses in individual videos which provide reliable labels while removing inconsistencies. Our pose embeddings capture visual characteristics of human pose that can boost existing supervised representations in human pose estimation and retrieval. We report quantitative and qualitative results on these tasks in Olympic Sports, Leeds Pose Sports and MPII Human Pose datasets.

****

Approximate Grassmannian Intersections: Subspace-Valued Subspace Learning

Calvin Murdock, Fernando De la Torre; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4308-4316

Subspace learning is one of the most foundational tasks in computer vision with applications ranging from dimensionality reduction to data denoising. As geometric objects, subspaces have also been successfully used for efficiently representing certain types of invariant data. However, methods for subspace learning from subspace-valued data have been notably absent due to incompatibilities with standard problem formulations. To fill this void, we introduce Approximate Grassmannian Intersections (AGI), a novel geometric interpretation of subspace learning posed as finding the approximate intersection of constraint sets on a Grassmann manifold. Our approach can naturally be applied to input subspaces of varying dimension while reducing to standard subspace learning in the case of vector-valued data. Despite the nonconvexity of our problem, its globally-optimal solution can be found using a singular value decomposition. Furthermore, we also propose an efficient, general optimization approach that can incorporate additional constraints to encourage properties such as robustness. Alongside standard subspace applications, AGI also enables the novel task of transfer learning via subspace completion. We evaluate our approach on a variety of applications, demonstrating improved invariance and generalization over vector-valued alternatives.

****

Side Information in Robust Principal Component Analysis: Algorithms and Applications

Niannan Xue, Yannis Panagakis, Stefanos Zafeiriou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4317-4325

Robust Principal Component Analysis (RPCA) aims at recovering a low-rank subspace from grossly corrupted high-dimensional (often visual) data and is a cornerstone in many machine learning and computer vision applications. Even though RPCA has been shown to be very successful in solving many rank minimisation problems, there are still cases where degenerate or suboptimal solutions are obtained. This is likely to be remedied by taking into account of domain-dependent prior knowledge. In this paper, we propose two models for the RPCA problem with the aid of side information on the low-rank structure of the data. The versatility of the proposed methods is demonstrated by applying them to four applications, namely background subtraction, facial image denoising, face and facial expression recognition. Experimental results on synthetic and five real world datasets indicate the robustness and effectiveness of the proposed methods on these application domains, largely outperforming six previous approaches.

****

Summarization and Classification of Wearable Camera Streams by Learning the Distributions Over Deep Features of Out-Of-Sample Image Sequences

Alessandro Perina, Sadegh Mohammadi, Nebojsa Jojic, Vittorio Murino; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4326-4334

A popular approach to training classifiers of new image classes is to use lower levels of a pre-trained feed-forward neural network and retrain only the top. Thus, most layers simply serve as highly nonlinear feature extractors. While these features were found useful for classifying a variety of scenes and objects, previous work also demonstrated unusual levels of sensitivity to the input especially for images which are veering too far away from the training distribution. This can lead to surprising results as an imperceptible change in an image can be enough to completely change the predicted class. This occurs in particular in applications involving personaldata, typically acquired with wearable cameras (e.g., visual lifelogs), where the problem is also made more complex by the dearth of new labeled training data that make supervised learning with deep models difficult. To alleviate these problems, in this paper we propose a new generative model that captures the feature distribution in new data. Its latent space then becomes more representative of the new data, while still retaining the generalization properties. In particular, we use constrained Markov walks over a counting grid for modeling image sequences, which not only yield good latent representations, but allow for excellent classification with only a handful of labeled training examples of the new scenes or objects, a scenario typical in lifelogging applications.

****

Unsupervised Learning From Video to Detect Foreground Objects in Single Images

Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4335-4343

Unsupervised learning from visual data is one of the most difficult challenges in computer vision. It is essential for understanding how visual recognition works. Learning from unsupervised input has an immense practical value, as huge quantities of unlabeled videos can be collected at low cost. Here we address the tas k of unsupervised learning to detect and segment foreground objects in single images. We achieve our goal by training a student pathway, consisting of a deep neural network that learns to predict, from a single input image, the output of a teacher pathway that performs unsupervised object discovery in video. Our approach is different from the published methods that perform unsupervised discovery in videos or in collections of images at test time. We move the unsupervised discovery phase during the training stage, while at test time we apply the standard feed-forward processing along the student pathway. This has a dual benefit: firstly, it allows, in principle, unlimited generalization possibilities during training, while remaining fast at testing. Secondly, the student not only becomes able to detect in single images significantly better than its unsupervised video discovery teacher, but it also achieves state of the art results on two current benchmarks, YouTube Objects and Object Discovery datasets. At test time, our system is two orders of magnitude faster than other previous methods.

****

Supplementary Meta-Learning: Towards a Dynamic Model for Deep Neural Networks

Feihu Zhang, Benjamin W. Wah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4344-4353

Data diversity in terms of types, styles, as well as radiometric, exposure and texture conditions widely exists in training and test data of vision applications. However, learning in traditional neural networks (NNs) only tries to find a model with fixed parameters that optimize the average behavior over all inputs, without using data-specific properties. In this paper, we develop a meta-level NN (MLNN) model that learns meta-knowledge on data-specific properties of images during learning and that dynamically adapts its weights during application according to the properties of the images input. MLNN consists of two parts: the dynamic supplementary NN (SNN) that learns meta-information on each type of inputs, and the fixed base-level NN (BLNN) that incorporates the meta-information from SNN into its weights at run time to realize the generalization for each type of inputs. We verify our approach using over ten network architectures under various application scenarios and loss functions. In low-level vision applications on image super-resolution and denoising, MLNN has 0.1 0.3 dB improvements on PSNR, whereas for high-level image classification, MLNN has accuracy improvement of 0.4 0.6% for Cifar10 and 1.2 2.1% for ImageNet when compared to convolutional NNs (CNNs). Improvements are more pronounced as the scale or diversity of data is increased.

****

Adversarial Inverse Graphics Networks: Learning 2D-To-3D Lifting and Image-To-Image Translation From Unpaired Supervision

Hsiao-Yu Fish Tung, Adam W. Harley, William Seto, Katerina Fragkiadaki; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4354-4362

Researchers have developed excellent feed-forward models that learn to map images to desired outputs, such as to the images' latent factors, or to other images, using supervised learning. Learning such mappings from unlabelled data, or improving upon supervised models by exploiting unlabelled data, remains elusive. We argue that there are two important parts to learning without annotations: (i) matching the predictions to the input observations, and (ii) matching the predictions to known priors. We propose Adversarial Inverse Graphics networks (AIGNs): weakly supervised neural network models that combine feedback from rendering their predictions, with distribution matching between their predictions and a collection of ground-truth factors. We apply AIGNs to 3D human pose estimation and 3D structure and egomotion estimation, and outperform models supervised by only paired annotations. We further apply AIGNs to facial image transformation using super-resolution and inpainting renderers, while deliberately adding biases in the ground-truth datasets. Our model seamlessly incorporates such biases, rendering input faces towards young, old, feminine, masculine or Tom Cruise-like equivalents (depending on the chosen bias), or adding lip and nose augmentations while inpainting concealed lips and noses.

****

Active Learning for Human Pose Estimation

Buyu Liu, Vittorio Ferrari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4363-4372

Annotating human poses in realistic scenes is very time consuming, yet necessary for training human pose estimators. We propose to address this problem in an active learning framework, which alternates between requesting the most useful annotations among a large set of unlabelled images, and re-training the pose estimator. To this end, (1) we propose an uncertainty estimator specific for body joint predictions, which takes into account the spatial distribution of the responses of the current pose estimator on the unlabelled images; (2) we propose a dynamic combination of influence and uncertainty cues, where their weights vary during the active learning process according to the reliability of the current pose estimator; (3) we introduce a computer assisted annotation interface, which reduces the time necessary for a human annotator to click on a joint by discretizing the image into regions generated by the current pose estimator. Experiments using the MPII and LSP datasets with both simulated and real annotators show that (1) the proposed active selection scheme outperforms several baselines; (2) our computer-assisted interface can further reduce annotation effort; and (3) our technique can further improve the performance of a pose estimator even when starting from an already strong one.performance in 23% annotation time.

****

Interleaved Group Convolutions

Ting Zhang, Guo-Jun Qi, Bin Xiao, Jingdong Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4373-4382

In this paper, we present a simple and modularized neural network architecture, named interleaved group convolutional neural networks (IGCNets). The main point lies in a novel building block, a pair of two successive interleaved group convolutions: primary group convolution and secondary group convolution. The two group convolutions are complementary: (i) the convolution on each partition in primary group convolution is a spatial convolution, while on each partition in secondary group convolution, the convolution is a point-wise convolution; (ii) the channels in the same secondary partition come from different primary partitions. We discuss one representative advantage: Wider than a regular convolution with the number of parameters and the computation complexity preserved. We also show that regular convolutions, group convolution with summation fusion, and the Xception block are special cases of interleaved group convolutions. Empirical results over standard benchmarks, CIFAR-10, CIFAR-100, SVHN and ImageNet demonstrate that our networks are more efficient in using parameters and computation complexity with similar or higher accuracy.

****

Learning-Based Cloth Material Recovery From Video

Shan Yang, Junbang Liang, Ming C. Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4383-4393

Image understanding enables better reconstruction of the physical world from images and videos. Existing methods focus largely on geometry and visual appearance of the reconstructed scene. In this paper, we extend the frontier in image understanding and present a new technique to recover the material properties of cloth from a video.Previous cloth material recovery methods often require markers or complex experimental set-up to acquire physical properties, or are limited to certain types of images/videos. Our approach takes advantages of the appearance changes of the moving cloth to infer its physical properties. To extract information about the cloth, our method characterizes both the motion space and the visual appearance of the cloth geometry. We apply the Convolutional Neural Network (CNN) and the Long Short Term Memory (LSTM) neural network to material recovery of cloth properties from videos. We also exploit simulated data to help statistical learning of mapping between the visual appearance and motion dynamics of the cloth. The effectiveness of our method is demonstrated via validation using simulated datasets and real-life recorded videos.

****

Unsupervised Video Understanding by Reconciliation of Posture Similarities

Timo Milbich, Miguel Bautista, Ekaterina Sutter, Bjorn Ommer; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4394-4404

Understanding human activity and being able to explain it in detail surpasses mere action classification by far in both complexity and value. The challenge is thus to describe an activity on the basis of its most fundamental constituents, the individual postures and their distinctive transitions. Supervised learning of such a fine-grained representation based on elementary poses is very tedious and does not scale. Therefore, we propose a completely unsupervised deep learning procedure based solely on video sequences, which starts from scratch without requiring pre-trained networks, predefined body models, or keypoints. A combinatorial sequence matching algorithm proposes relations between frames from subsets of the training data, while a CNN is reconciling the transitivity conflicts of the different subsets to learn a single concerted pose embedding despite changes in appearance across sequences. Without any manual annotation, the model learns a structured representation of postures and their temporal development. The model not only enables retrieval of similar postures but also temporal super-resolution. Additionally, based on a recurrent formulation, next frames can be synthesized.

****

Action Tubelet Detector for Spatio-Temporal Action Localization

Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4405-4413

Current state-of-the-art approaches for spatio-temporal action localization rely on detections at the frame level that are then linked or tracked across time. In this paper, we leverage the temporal continuity of videos instead of operating at the frame level. We propose the ACtion Tubelet detector (ACT-detector) that takes as input a sequence of frames and outputs tubelets, ie, sequences of bounding boxes with associated scores. The same way state-of-the-art object detectors rely on anchor boxes, our ACT-detector is based on anchor cuboids. We build upon the SSD framework. Convolutional features are extracted for each frame, while scores and regressions are based on the temporal stacking of these features, thus exploiting information from a sequence. Our experimental results show that leveraging sequences of frames significantly improves detection performance over using individual frames. The gain of our tubelet detector can be explained by both more accurate scores and more precise localization. Our ACT-detector outperforms the state-of-the-art methods for frame-mAP and video-mAP on the J-HMDB and UCF-101 datasets, in particular at high overlap thresholds.

****

AMTnet: Action-Micro-Tube Regression by End-To-End Trainable Deep Architecture

Suman Saha, Gurkirt Singh, Fabio Cuzzolin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4414-4423

Dominant approaches to action detection can only provide sub-optimal solutions to the problem, as they rely on seeking frame-level detections, to later compose them into "action tubes" in a post-processing step. With this paper we radically depart from current practice, and take a first step towards the design and implementation of a deep network architecture able to classify and regress whole video subsets, so providing a truly optimal solution of the action detection problem. In this work, in particular, we propose a novel deep net framework able to regress and classify 3D region proposals spanning two successive video frames, whose core is an evolution of classical region proposal networks (RPNs). As such, our 3D-RPN net is able to effectively encode the temporal aspect of actions by purely exploiting appearance, as opposed to methods which heavily rely on expensive flow maps. The proposed model is end-to-end trainable and can be jointly optimised for action localisation and classification in a single step. At test time the network predicts "micro-tubes" encompassing two successive frames, which are linked up into complete action tubes via a new algorithm which exploits the temporal encoding learned by the network and cuts computation time by 50%. Promising results on the J-HMDB-21 and UCF-101 action detection datasets show that our model does outperform the state-of-the-art when relying purely on appearance.

****

Constrained Convolutional Sparse Coding for Parametric Based Reconstruction of Line Drawings

Sara Shaheen, Lama Affara, Bernard Ghanem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4424-4432

Convolutional sparse coding (CSC) plays an essential role in many computer vision applications ranging from image compression to deep learning. In this work, we spot the light on a new application where CSC can effectively serve, namely line drawing analysis. The process of drawing a line drawing can be approximated as the sparse spatial localization of a number of typical basic strokes, which in turn can be cast as a non-standard CSC model that considers the line drawing formation process from parametric curves. These curves are learned to optimize the fit between the model and a specific set of line drawings. Parametric representation of sketches is vital in enabling automatic sketch analysis, synthesis and manipulation. A couple of sketch manipulation examples are demonstrated in this work. Consequently, our novel method is expected to provide a reliable and automatic method for parametric sketch description. Through experiments, we empirically validate the convergence of our method to a feasible solution.

****

Neural Ctrl-F: Segmentation-Free Query-By-String Word Spotting in Handwritten Manuscript Collections

Tomas Wilkinson, Jonas Lindstrom, Anders Brun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4433-4442

In this paper, we approach the problem of segmentation-free query-by-string word spotting for handwritten documents. In other words, we use methods inspired from computer vision and machine learning to search for words in large collections of digitized manuscripts. In particular, we are interested in historical handwritten texts, which are often far more challenging than modern printed documents. This task is important, as it provides people with a way to quickly find what they are looking for in large collections that are tedious and difficult to read manually. To this end, we introduce an end-to-end trainable model based on deep neural networks that we call Ctrl-F-Net. Given a full manuscript page, the model simultaneously generates region proposals, and embeds these into a distributed word embedding space, where searches are performed. We evaluate the model on common benchmarks for handwritten word spotting, outperforming the previous state-of-the-art segmentation-free approaches by a large margin, and in some cases even segmentation-based approaches. One interesting real-life application of our approach is to help historians to find and count specific words in court records that are related to women's sustenance activities and division of labor. We provide promising preliminary experiments that validate our method on this task.

****

Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions

Pascal Mettes, Cees G. M. Snoek; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4443-4452

We aim for zero-shot localization and classification of human actions in video. Where traditional approaches rely on global attribute or object classification scores for their zero-shot knowledge transfer, our main contribution is a spatial-aware object embedding. To arrive at spatial awareness, we build our embedding on top of freely available actor and object detectors. Relevance of objects is determined in a word embedding space and further enforced with estimated spatial preferences. Besides local object awareness, we also embed global object awareness into our embedding to maximize actor and object interaction. Finally, we exploit the object positions and sizes in the spatial-aware embedding to demonstrate a new spatio-temporal action retrieval scenario with composite queries. Action localization and classification experiments on four contemporary action video datasets support our proposal. Apart from state-of-the-art results in the zero-shot localization and classification settings, our spatial-aware embedding is even competitive with recent supervised action localization alternatives.

****

Semantic Video CNNs Through Representation Warping
Raghudeep Gadde, Varun Jampani, Peter V. Gehler; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4453-4462

In this work, we propose a technique to convert CNN models for semantic segmentation of static images into CNNs for video data. We describe a warping method that can be used to augment existing architectures with very little extra computational cost. This module is called NetWarp and we demonstrate its use for a range of network architectures. The main design principle is to use optical flow of adjacent frames for warping internal network representations across time. A key insight of this work is that fast optical flow methods can be combined with many different CNN architectures for improved performance and end-to-end training. Experiments validate that the proposed approach incurs only little extra computational cost, while improving performance, when video streams are available. We achieve new state-of-the-art results on the CamVid and Cityscapes benchmark datasets and show consistent improvements over different baseline networks. Our code and models are available at http://segmentation.is.tue.mpg.de

****

Video Frame Synthesis Using Deep Voxel Flow
Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, Aseem Agarwala; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4463-4471

We address the problem of synthesizing new video frames in an existing video, either in-between existing frames (interpolation), or subsequent to them (extrapolation). This problem is challenging because video appearance and motion can be highly complex. Traditional optical-flow-based solutions often fail where flow estimation is challenging, while newer neural-network-based methods that hallucinate pixel values directly often produce blurry results. We combine the advantages of these two methods by training a deep network that learns to synthesize video frames by flowing pixel values from existing ones, which we call deep voxel flow. Our method requires no human supervision, and any video can be used as training data by dropping, and then learning to predict, existing frames. The technique is efficient, and can be applied at any video resolution. We demonstrate that our method produces results that both quantitatively and qualitatively improve upon the state-of-the-art.

****

Detail-Revealing Deep Video Super-Resolution
Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4472-4480

Previous CNN-based video super-resolution approaches need to align multiple frames to the reference. In this paper, we show that proper frame alignment and motion compensation is crucial for achieving high quality results. We accordingly propose a 'sub-pixel motion compensation' (SPMC) layer in a CNN framework. Analysis and experiments show the suitability of this layer in video SR. The final end-to-end, scalable CNN framework effectively incorporates the SPMC layer and fuses multiple frames to reveal image details. Our implementation can generate visually and quantitatively high-quality results, superior to current state-of-the-arts, without the need of parameter tuning.

****

Learning Video Object Segmentation With Visual Memory
Pavel Tokmakov, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4481-4490

This paper addresses the task of segmenting moving objects in unconstrained videos. We introduce a novel two-stream neural network with an explicit memory module to achieve this. The two streams of the network encode spatial and temporal features in a video sequence respectively, while the memory module captures the evolution of objects over time. The module to build a 'visual memory' in video, i.e., a joint representation of all the video frames, is realized with a convolutional recurrent unit learned from a small number of training video sequences. Given a video frame as input, our approach assigns each pixel an object or background label based on the learned spatio-temporal features as well as the 'visual me mory' specific to the video, acquired automatically without any manually-annotated frames. We evaluate our method extensively on two benchmarks, DAVIS and Freiburg-Berkeley motion segmentation datasets, and show state-of-the-art results. For example, our approach outperforms the top method on the DAVIS dataset by nearly 6%. We also provide an extensive ablative analysis to investigate the influence of each component in the proposed framework.

****

EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis

Mehdi S. M. Sajjadi, Bernhard Scholkopf, Michael Hirsch; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4491-4500

Single image super-resolution is the task of inferring a high-resolution image from a single low-resolution input. Traditionally, the performance of algorithms for this task is measured using pixel-wise reconstruction measures such as peak signal-to-noise ratio (PSNR) which have been shown to correlate poorly with the human perception of image quality. As a result, algorithms minimizing these metrics tend to produce over-smoothed images that lack high-frequency textures and do not look natural despite yielding high PSNR values. We propose a novel application of automated texture synthesis in combination with a perceptual loss focusing on creating realistic textures rather than optimizing for a pixel-accurate reproduction of ground truth images during training. By using feed-forward fully convolutional neural networks in an adversarial training setting, we achieve a significant boost in image quality at high magnification ratios. Extensive experiments on a number of datasets show the effectiveness of our approach, yielding state-of-the-art results in both quantitative and qualitative benchmarks.

****

Makeup-Go: Blind Reversion of Portrait Edit

Ying-Cong Chen, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4501-4509

Virtual face beautification (or markup) becomes common operations in camera or image processing Apps, which is actually deceiving. In this paper, we propose the task of restoring a portrait image from this process. As the first attempt along this line, we assume unknown global operations on human faces and aim to tackle the two issues of skin smoothing and skin color change. These two tasks, intriguingly, impose very different difficulties to estimate subtle details and major color variation. We propose a Component Regression Network (CRN) and address the limitation of using Euclidean loss in blind reversion. CRN maps the edited portrait images back to the original ones without knowing beautification operation details. Our experiments demonstrate effectiveness of the system for this novel task.

****

Shadow Detection With Conditional Generative Adversarial Networks

Vu Nguyen, Tomas F. Yago Vicente, Maozheng Zhao, Minh Hoai, Dimitris Samaras; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4510-4518

We introduce scGAN, a novel extension of conditional Generative Adversarial Networks (GAN) tailored for the challenging problem of shadow detection in images. Previous methods for shadow detection focus on learning the local appearance of shadow regions, while using limited local context reasoning in the form of pairwise potentials in a Conditional Random Field. In contrast, the proposed adversarial approach is able to model higher level relationships and global scene characteristics. We train a shadow detector that corresponds to the generator of a conditional GAN, and augment its shadow accuracy by combining the typical GAN loss with a data loss term. Due to the unbalanced distribution of the shadow labels, we use weighted cross entropy. With the standard GAN architecture, properly setting the weight for the cross entropy would require training multiple GANs, a computationally expensive grid procedure. In scGAN, we introduce an additional sensitivity parameter w to the generator. The proposed approach effectively parameterizes the loss of the trained detector. The resulting shadow detector is a single network that can generate shadow maps corresponding to different sensitivity levels, obviating the need for multiple models and a costly training procedure. We evaluate our method on the large-scale SBU and UCF shadow datasets, and observe up to 17% error reduction with respect to the previous state-of-the-art method.

****

Learning High Dynamic Range From Outdoor Panoramas

Jinsong Zhang, Jean-Francois Lalonde; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4519-4528

Outdoor lighting has extremely high dynamic range. This makes the process of capturing outdoor environment maps notoriously challenging since special equipment must be used. In this work, we propose an alternative approach. We first capture lighting with a regular, LDR omnidirectional camera, and aim to recover the HDR after the fact via a novel, learning-based inverse tonemapping method. We propose a deep autoencoder framework which regresses linear, high dynamic range data from non-linear, saturated, low dynamic range panoramas. We validate our method through a wide set of experiments on synthetic data, as well as on a novel dataset of real photographs with ground truth. Our approach finds applications in a variety of settings, ranging from outdoor light capture to image matching.

****

DCTM: Discrete-Continuous Transformation Matching for Semantic Flow

Seungryong Kim, Dongbo Min, Stephen Lin, Kwanghoon Sohn; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4529-4538

Techniques for dense semantic correspondence have provided limited ability to deal with the geometric variations that commonly exist between semantically similar images. While variations due to scale and rotation have been examined, there is a lack of practical solutions for more complex deformations such as affine transformations because of the tremendous size of the associated solution space. To address this problem, we present a discrete-continuous transformation matching (DCTM) framework where dense affine transformation fields are inferred through a discrete label optimization in which the labels are iteratively updated via continuous regularization. In this way, our approach draws solutions from the continuous space of affine transformations in a manner that can be computed efficiently through constant-time edge-aware filtering and a proposed affine-varying CNN-based descriptor. Experimental results show that this model outperforms the state-of-the-art methods for dense semantic correspondence on various benchmarks.

****

MemNet: A Persistent Memory Network for Image Restoration

Ying Tai, Jian Yang, Xiaoming Liu, Chunyan Xu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4539-4547

Recently, very deep convolutional neural networks (CNNs) have been attracting considerable attention in image restoration. However, as the depth grows, the long-term dependency problem is rarely realized for these very deep models, which results in the prior states/layers having little influence on the subsequent ones. Motivated by the fact that human thoughts have persistency, we propose a very deep persistent memory network (MemNet) that introduces a memory block, consisting of a recursive unit and a gate unit, to explicitly mine persistent memory through an adaptive learning process. The recursive unit learns multi-level representations of the current state under different receptive fields. The representations and the outputs from the previous memory blocks are concatenated and sent to the gate unit, which adaptively controls how much of the previous states should be reserved, and decides how much of the current state should be stored. We apply MemNet to three image restoration tasks, i.e., image denosing, super-resolution and JPEG deblocking. Comprehensive experiments demonstrate the necessity of the MemNet and its unanimous superiority on all three tasks over the state of the arts. Code is available at https://github.com/tyshiwo/MemNet.

****

Structure-Measure: A New Way to Evaluate Foreground Maps

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, Ali Borji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4548-4557

Foreground map evaluation is crucial for gauging the progress of object segmentation algorithms, in particular in the filed of salient object detection where the purpose is to accurately detect and segment the most salient object in a scene

. Several widely-used measures such as Area Under the Curve (AUC), Average Precision (AP) and the recently proposed Fbw have been utilized to evaluate the similarity between a non-binary saliency map (SM) and a ground-truth (GT) map. These measures are based on pixel-wise errors and often ignore the structural similarities. Behavioral vision studies, however, have shown that the human visual system is highly sensitive to structures in scenes. Here, we propose a novel, efficient, and easy to calculate measure known an structural similarity measure (Structure-measure) to evaluate non-binary foreground maps. Our new measure simultaneously evaluates region-aware and object-aware structural similarity between a SM and a GT map. We demonstrate superiority of our measure over existing ones using 5 meta-measures on 5 benchmark datasets.
****

Weakly- and Self-Supervised Learning for Content-Aware Deep Image Retargeting
Donghyeon Cho, Jinsun Park, Tae-Hyun Oh, Yu-Wing Tai, In So Kweon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4558-4567

This paper proposes a weakly- and self-supervised deep convolutional neural network (WSSDCNN) for content-aware image retargeting. Our network takes a source image and a target aspect ratio, and then directly outputs a retargeted image. Retargeting is performed through a shift map, which is a pixel-wise mapping from the source to the target grid. Our method implicitly learns an attention map, which leads to a content-aware shift map for image retargeting. As a result, discriminative parts in an image are preserved, while background regions are adjusted seamlessly. In the training phase, pairs of an image and its image level annotation are used to compute content and structure losses. We demonstrate the effectiveness of our proposed method for a retargeting application with insightful analyses.
****

Practical and Efficient Multi-View Matching
Eleonora Maset, Federica Arrigoni, Andrea Fusiello; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4568-4576

In this paper we propose a novel solution to the multi-view matching problem that, given a set of noisy pairwise correspondences, jointly updates them so as to maximize their consistency. Our method is based on a spectral decomposition, resulting in a closed-form efficient algorithm, in contrast to other iterative techniques that can be found in the literature. Experiments on both synthetic and real datasets show that our method achieves comparable or superior accuracy to state-of-the-art algorithms in significantly less time. We also demonstrate that our solution can efficiently handle datasets of hundreds of images, which is unprecedented in the literature.
****

Unrolled Memory Inner-Products: An Abstract GPU Operator for Efficient Vision-Related Computations
Yu-Sheng Lin, Wei-Chao Chen, Shao-Yi Chien; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4577-4585

Recently, convolutional neural networks (CNNs) have achieved great success in fields such as computer vision, natural language processing, and artificial intelligence. Many of these applications utilize parallel processing in GPUs to achieve higher performance. However, it remains a daunting task to optimize for GPUs, and most researchers have to rely on vendor-provided libraries for such purposes. In this paper, we discuss an operator that can be used to succinctly express computational kernels in CNNs and various scientific and vision applications. This operator, called Unrolled-Memory-Inner-Product (UMI), is a computationally-efficient operator with smaller code token requirement. Since a naive UMI implementation would increase memory requirement through input data unrolling, we propose a method to achieve optimal memory fetch performance in modern GPUs. We demonstrate this operator by converting several popular applications into the UMI representation and achieve 1.3x-26.4x speedup against frameworks such as OpenCV and Caffe.
****

Learning to Push the Limits of Efficient FFT-Based Image Deconvolution

Jakob Kruse, Carsten Rother, Uwe Schmidt; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4586-4594

This work addresses the task of non-blind image deconvolution. Motivated to keep up with the constant increase in image size, with megapixel images becoming the norm, we aim at pushing the limits of efficient FFT-based techniques. Based on an analysis of traditional and more recent learning-based methods, we generalize existing discriminative approaches by using more powerful regularization, based on convolutional neural networks. Additionally, we propose a simple, yet effective, boundary adjustment method that alleviates the problematic circular convolution assumption, which is necessary for FFT-based deconvolution. We evaluate our approach on two common non-blind deconvolution benchmarks and achieve state-of-the-art results even when including methods which are computationally considerably more expensive.

****

Learning Spread-Out Local Feature Descriptors

Xu Zhang, Felix X. Yu, Sanjiv Kumar, Shih-Fu Chang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4595-4603

We propose a simple, yet powerful regularization technique that can be used to significantly improve both the pairwise and triplet losses in learning local feature descriptors. The idea is that in order to fully utilize the expressive power of the descriptor space, good local feature descriptors should be sufficiently "spread-out" over the space. In this work, we propose a regularization term to maximize the spread in feature descriptor inspired by the property of uniform distribution. We show that the proposed regularization with triplet loss outperforms existing Euclidean distance based descriptor learning techniques by a large margin. As an extension, the proposed regularization technique can also be used to improve image-level deep feature embedding.

****

Visual Odometry for Pixel Processor Arrays

Laurie Bose, Jianing Chen, Stephen J. Carey, Piotr Dudek, Walterio Mayol-Cuevas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4604-4612

We present an approach of estimating constrained motion of a novel Cellular Processor Array (CPA) camera, on which each pixel is capable of limited processing and data storage allowing for fast low power parallel computation to be carried out directly on the focal-plane of the device. Rather than the standard pipeline involved with traditional cameras whereby whole camera images are transferred to a general computer system for processing, our approach performs all computation upon the CPA itself, with the only information being transfered to a standard computer being the camera's estimated motion.This limited data transfer allows for high frame-rate processing at hundreds of hz while consuming less than 1.5 Watts of power.The current implementation is restricted to the estimation of the camera's rotation in yaw and pitch, along with a scaleless estimate of the camera's forward and backward translation. We describe methods of image alignment by gradient descent, edge detection, and image scaling, all of which are performed solely on the CPA device itself and which form the core components of detecting camera motion.

****

Joint Estimation of Camera Pose, Depth, Deblurring, and Super-Resolution From a Blurred Image Sequence

Haesol Park, Kyoung Mu Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4613-4621

The conventional methods for estimating camera poses and scene structures from severely blurry or low resolution images often result in failure. The off-the-shelf deblurring or super resolution methods may show visually pleasing results. However, applying each technique independently before matching is generally unprofitable because this naive series of procedures ignores the consistency between images. In this paper, we propose a pioneering unified framework that solves four problems simultaneously, namely, dense depth reconstruction, camera pose estima tion, super resolution, and deblurring. By reflecting a physical imaging process, we formulate a cost minimization problem and solve it using an alternating optimization technique. The experimental results on both synthetic and real videos show high-quality depth maps derived from severely degraded images that contrast the failures of naive multi-view stereo methods. Our proposed method also produces outstanding deblurred and super-resolved images unlike the independent application or combination of conventional video deblurring, super resolution methods.

****

2D-Driven 3D Object Detection in RGB-D Images

Jean Lahoud, Bernard Ghanem; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4622-4630

In this paper, we present a technique that places 3D bounding boxes around objects in an RGB-D scene. Our approach makes best use of the 2D information to quickly reduce the search space in 3D, benefiting from state-of-the-art 2D object detection techniques. We then use the 3D information to orient, place, and score bounding boxes around objects. We independently estimate the orientation for every object, using previous techniques that utilize normal information. Object locations and sizes in 3D are learned using a multilayer perceptron (MLP). In the final step, we refine our detections based on object class relations within a scene. When compared to state-of-the-art detection methods that operate almost entirely in the sparse 3D domain, extensive experiments on the well-known SUN RGB-D dataset show that our proposed method is much faster (4.1s per image) in detecting 3D objects in RGB-D images and performs better (3 mAP higher) than the state-of-the-art method that is 4.7 times slower and comparably to the method that is two orders of magnitude slower. This work hints at the idea that 2D-driven object detection in 3D should be further explored, especially in cases where the 3D input is sparse.

****

Ray Space Features for Plenoptic Structure-From-Motion

Yingliang Zhang, Peihong Yu, Wei Yang, Yuanxi Ma, Jingyi Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4631-4639

Traditional Structure-from-Motion (SfM) uses images captured by cameras as inputs. In this paper, we explore using light fields captured by plenoptic cameras or camera arrays as inputs. We call this solution plenoptic SfM or P-SfM solution. We first present a comprehensive theory on ray geometry transforms under light field pose variations. We derive the transforms of three typical ray manifolds: rays passing through a point or point-ray manifold, rays passing through a 3D line or ray-line manifold, and rays lying on a common 3D plane or ray-plane manifold. We show that by matching these manifolds across LFs, we can recover light field poses and conduct bundle adjustment in ray space. We validate our theory and framework on synthetic and real data on light fields of different scales: small scale LFs acquired using a LF camera and large scale LFs by a camera array. We show that our P-SfM technique can significantly improve the accuracy and reliability over regular SfM and PnP especially on traditionally challenging scenes where reliable feature point correspondences are difficult to obtain but line or plane correspondences are readily accessible.

****

Depth Estimation Using Structured Light Flow -- Analysis of Projected Pattern Flow on an Object's Surface

Ryo Furukawa, Ryusuke Sagawa, Hiroshi Kawasaki; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4640-4648

Shape reconstruction techniques using structured light have been widely researched and developed due to their robustness, high precision, and density. Because the techniques are based on decoding a pattern to find correspondences, it implicitly requires that the projected patterns be clearly captured by an image sensor, i.e., to avoid defocus and motion blur of the projected pattern. Although intensive researches have been conducted for solving defocus blur, few researches for motion blur and only solution is to capture with extremely fast shutter speed. In this paper, unlike the previous approaches, we actively utilize motion blur, which we refer to as a light flow, to estimate depth. Analysis reveals that minimum two light flows, which are retrieved from two projected patterns on the object, are required for depth estimation. To retrieve two light flows at the same time, two sets of parallel line patterns are illuminated from two video projectors and the size of motion blur of each line is precisely measured. By analyzing the light flows, i.e. lengths of the blurs, scene depth information is estimated. In the experiments, 3D shapes of fast moving objects, which are inevitably captured with motion blur, are successfully reconstructed by our technique.

****

Monocular Dense 3D Reconstruction of a Complex Dynamic Scene From Two Perspective Frames

Suryansh Kumar, Yuchao Dai, Hongdong Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4649-4657

This paper proposes a new approach for monocular dense 3D reconstruction of a complex dynamic scene from two perspective frames. By applying superpixel oversegmentation to the image, we model a generically dynamic (hence non-rigid) scene with a piecewise planar and rigid approximation. In this way, we reduce the dynamic reconstruction problem to a "3D jigsaw puzzle" problem which takes pieces from an unorganized "soup of superpixels". We show that our method provides an effective solution to the inherent relative scale ambiguity in structure-from-motion. Since our method does not assume a template prior, or per-object segmentation, or knowledge about the rigidity of the dynamic scene, it is applicable to a wide range of scenarios. Extensive experiments on both synthetic and real monocular sequences demonstrate the superiority of our method compared with the state-of-the-art methods.

****

Optimal Transformation Estimation With Semantic Cues

Danda Pani Paudel, Adlane Habed, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4658-4667

This paper addresses the problem of estimating the geometric transformation relating two distinct visual modalities (e.g. an image and a map, or a projective structure and a Euclidean 3D model) while relying only on semantic cues, such as semantically segmented regions or object bounding boxes. The proposed approach differs from the traditional feature-to-feature correspondence reasoning: starting from semantic regions on one side, we seek their possible corresponding regions on the other, thus constraining the sought geometric transformation. This entails a simultaneous search for the transformation and for the region-to-region correspondences.This paper is the first to derive the conditions that must be satisfied for a convex region, defined by control points, to be transformed inside an ellipsoid. These conditions are formulated as Linear Matrix Inequalities and used within a Branch-and-Prune search to obtain the globally optimal transformation. We tested our approach, under mild initial bound conditions, on two challenging registration problems for aligning: (i) a semantically segmented image and a map via a 2D homography; (ii) a projective 3D structure and its Euclidean counterpart.

****

Dynamics Enhanced Multi-Camera Motion Segmentation From Unsynchronized Videos

Xikang Zhang, Bengisu Ozbay, Mario Sznaier, Octavia Camps; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4668-4676

This paper considers the multi-camera motion segmentation problem using unsynchronized videos. Specifically, given two video clips containing several moving objects, captured by unregistered, unsynchronized cameras with different viewpoints, our goal is to assign features to moving objects in the scene. This problem challenges existing methods, due to the lack of registration information and correspondences across cameras. To solve it, we propose a new method that exploits both shape and dynamical information and does not require spatio-temporal registration or shared features. As shown in the paper, the combination of shape and dynamical information results in improved performance even in the single camera case, and allows for solving the multi-camera segmentation problem with a computational cost similar to that of existing single-view techniques. These results are illustrated using both the existing Hopkins 155 data set and a new multi-camera data set, the RSL-12.

****

Taking the Scenic Route to 3D: Optimising Reconstruction From Moving Cameras

Oscar Mendez, Simon Hadfield, Nicolas Pugeault, Richard Bowden; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4677-4685

Reconstruction of 3D environments is a problem that has been widely addressed in the literature. While many approaches exist to perform reconstruction, few of them take an active role in deciding where the next observations should come from. Furthermore, the problem of travelling from the camera's current position to the next, known as pathplanning, usually focuses on minimising path length. This approach is ill-suited for reconstruction applications, where learning about the environment is more valuable than speed of traversal. We present a novel Scenic Route Planner that selects paths which maximise information gain, both in terms of total map coverage and reconstruction accuracy. We also introduce a new type of collaborative behaviour into the planning stage called opportunistic collaboration, which allows sensors to switch between acting as independent Structure from Motion (SfM) agents or as a variable baseline stereo pair. We show that Scenic Planning enables similar performance to state-of-the-art batch approaches using less than 0.00027% of the possible stereo pairs (3% of the views). Comparison against length-based pathplanning approaches show that our approach produces more complete and more accurate maps with fewer frames. Finally, we demonstrate the Scenic Pathplanner's ability to generalise to live scenarios by mounting cameras on autonomous ground-based sensor platforms and exploring an environment.

****

FLaME: Fast Lightweight Mesh Estimation Using Variational Smoothing on Delaunay Graphs

W. Nicholas Greene, Nicholas Roy; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4686-4694

We propose a lightweight method for dense online monocular depth estimation capable of reconstructing 3D meshes on computationally constrained platforms. Our main contribution is to pose the reconstruction problem as a non-local variational optimization over a time-varying Delaunay graph of the scene geometry, which allows for an efficient, keyframeless approach to depth estimation. The graph can be tuned to favor reconstruction quality or speed and is continuously smoothed and augmented as the camera explores the scene. Unlike keyframe-based approaches, the optimized surface is always available at the current pose, which is necessary for low-latency obstacle avoidance. FLaME (Fast Lightweight Mesh Estimation) can generate mesh reconstructions at upwards of 230 Hz using less than one Intel i7 CPU core, which enables operation on size, weight, and power-constrained platforms. We present results from both benchmark datasets and experiments running FLaME in-the-loop onboard a small flying quadrotor.

****

Efficient Algorithms for Moral Lineage Tracing

Markus Rempfler, Jan-Hendrik Lange, Florian Jug, Corinna Blasse, Eugene W. Myers, Bjoern H. Menze, Bjoern Andres; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4695-4704

Lineage tracing, the joint segmentation and tracking of living cells as they move and divide in a sequence of light microscopy images, is a challenging task. Jug et al. have proposed a mathematical abstraction of this task, the moral lineage tracing problem (MLTP), whose feasible solutions define both a segmentation of every image and a lineage forest of cells. Their branch-and-cut algorithm, however, is prone to many cuts and slow convergence for large instances. To address this problem, we make three contributions: (i) we devise the first efficient primal feasible local search algorithms for the MLTP, (ii) we improve the branch-and-cut algorithm by separating tighter cutting planes and by incorporating our primal algorithms, (iii) we show in experiments that our algorithms find accurate solutions on the problem instances of Jug et al. and scale to larger instances, leveraging moral lineage tracing to practical significance.

****

From RGB to Spectrum for Natural Scenes via Manifold-Based Mapping
Yan Jia, Yinqiang Zheng, Lin Gu, Art Subpa-Asa, Antony Lam, Yoichi Sato, Imari Sato; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4705-4713

Spectral analysis of natural scenes can provide much more detailed information about the scene than an ordinary RGB camera. The richer information provided by hyperspectral images has been beneficial to numerous applications, such as understanding natural environmental changes and classifying plants and soils in agriculture based on their spectral properties. In this paper, we present an efficient manifold learning based method for accurately reconstructing a hyperspectral image from a single RGB image captured by a commercial camera with known spectral response. By applying a nonlinear dimensionality reduction technique to a large set of natural spectra, we show that the spectra of natural scenes lie on an intrinsically low dimensional manifold. This allows us to map an RGB vector to its corresponding hyperspectral vector accurately via our proposed novel manifold-based reconstruction pipeline. Experiments using both synthesized RGB images using hyperspectral datasets and real world data demonstrate our method outperforms the state-of-the-art.

****

DeepFuse: A Deep Unsupervised Approach for Exposure Fusion With Extreme Exposure Image Pairs
K. Ram Prabhakar, V Sai Srikar, R. Venkatesh Babu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4714-4722

We present a novel deep learning architecture for fusing static multi-exposure images. Current multi-exposure fusion (MEF) approaches use hand-crafted features to fuse input sequence. However, the weak hand-crafted representations are not robust to varying input conditions. Moreover, they perform poorly for extreme exposure image pairs. Thus, it is highly desirable to have a method that is robust to varying input conditions and capable of handling extreme exposure without artifacts. Deep representations have known to be robust to input conditions and have shown phenomenal performance in a supervised setting. However, the stumbling block in using deep learning for MEF was the lack of sufficient training data and an oracle to provide the ground-truth for supervision. To address the above issues, we have gathered a large dataset of multi-exposure image stacks for training and to circumvent the need for ground truth images, we propose an unsupervised deep learning framework for MEF utilizing a no-reference quality metric as loss function. The proposed approach uses a novel CNN architecture trained to learn the fusion operation without reference ground truth image. The model fuses a set of common low level features extracted from each image to generate artifact-free perceptually pleasing results. We perform extensive quantitative and qualitative evaluation and show that the proposed technique outperforms existing state-of-the-art approaches for a variety of natural images.

****

Learning Dense Facial Correspondences in Unconstrained Images
Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, Hao Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4723-4732

We present a minimalistic but effective neural network that computes dense facial correspondences in highly unconstrained RGB images. Our network learns a per-pixel flow and a matchability mask between 2D input photographs of a person and the projection of a textured 3D face model. To train such a network, we generate a massive dataset of synthetic faces with dense labels using renderings of a morphable face model with variations in pose, expressions, lighting, and occlusions. We found that a training refinement using real photographs is required to drastically improve the ability to handle real images. When combined with a facial detection and 3D face fitting step, we show that our approach outperforms the state-of-the-art face alignment methods in terms of accuracy and speed. By directly estimating dense correspondences, we do not rely on the full visibility of sparse facial landmarks and are not limited to the model space of regression-based approaches. We also assess our method on video frames and demonstrate successful per-frame processing under extreme pose variations, occlusions, and lighting con

ditions. Compared to existing 3D facial tracking techniques, our fitting does not rely on previous frames or frontal facial initialization and is robust to imperfect face detections.
****
Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-Identification
Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, Pan Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4733-4742
Person Re-Identification (person re-id) is a crucial task as its applications in visual surveillance and human-computer interaction. In this work, we present a novel joint Spatial and Temporal Attention Pooling Network (ASTPN) for video-based person re-identification, which enables the feature extractor to be aware of the current input video sequences, in a way that interdependency from the matching items can directly influence the computation of each other's representation. Specifically, the spatial pooling layer is able to select regions from each frame, while the attention temporal pooling performed can select informative frames over the sequence, both pooling guided by the information from distance matching. Experiments are conduced on the iLIDS-VID, PRID-2011 and MARS datasets and the results demonstrate that this approach outperforms existing state-of-art methods. We also analyze how the joint pooling in both dimensions can boost the person re-id performance more effectively than using either of them separately.
****
Automatic Content-Aware Projection for 360deg Videos
Yeong Won Kim, Chang-Ryeol Lee, Dae-Yong Cho, Yong Hoon Kwon, Hyeok-Jae Choi, Kuk-Jin Yoon; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4743-4751
To watch 360 videos on normal 2D displays, we need to project the selected part of the 360 image onto the 2D display plane. In this paper, we propose a fully-automated framework for generating content-aware 2D normal-view perspective videos from 360 videos. Especially, we focus on the projection step preserving important image contents and reducing image distortion. Basically, our projection method is based on Pannini projection model. At first, the salient contents such as linear structures and salient regions in the image are preserved by optimizing the single Panini projection model. Then, the multiple Panini projection models at salient regions are interpolated to suppress image distortion globally. Finally, the temporal consistency for image projection is enforced for producing temporally stable normal-view videos. Our proposed projection method does not require any user-interaction and is much faster than previous content-preserving methods. It can be applied to not only images but also videos taking the temporal consistency of projection into account. Experiments on various 360 videos show the superiority of the proposed projection method quantitatively and qualitatively.
****
Blur-Invariant Deep Learning for Blind-Deblurring
T. M. Nimisha, Akash Kumar Singh, A. N. Rajagopalan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4752-4760
In this paper, we investigate deep neural networks for blind motion deblurring. Instead of regressing for the motion blur kernel and performing non-blind deblurring out- side of the network (as most methods do), we propose a compact and elegant end-to-end deblurring network. Inspired by the data-driven sparse-coding approaches that are capable of capturing linear dependencies in data, we generalize this notion by embedding non-linearities into the learning process. We propose a new architecture for blind motion deblurring that consists of an autoencoder that learns the data prior, and an adversarial network that attempts to generate and discriminate between clean and blurred features. Once the network is trained, the generator learns a blur-invariant data representation which when fed through the decoder results in the final deblurred output.
****
Non-Linear Convolution Filters for CNN-Based Learning
Georgios Zoumpourlis, Alexandros Doumanoglou, Nicholas Vretos, Petros Daras; Pro

ceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4761-4769

During the last years, Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in image classification. Their architectures have largely drawn inspiration by models of the primate visual system. However, while recent research results of neuroscience prove the existence of non-linear operations in the response of complex visual cells, little effort has been devoted to extend the convolution technique to non-linear forms. Typical convolutional layers are linear systems, hence their expressiveness is limited. To overcome this, various non-linearities have been used as activation functions inside CNNs, while also many pooling strategies have been applied. We address the issue of developing a convolution method in the context of a computational model of the visual cortex, exploring quadratic forms through the Volterra kernels. Such forms, constituting a more rich function space, are used as approximations of the response profile of visual cells. Our proposed second-order convolution is tested on CIFAR-10 and CIFAR-100. We show that a network which combines linear and non-linear filters in its convolutional layers, can outperform networks that use standard linear filters with the same architecture, yielding results competitive with the state-of-the-art on these datasets.

****

AOD-Net: All-In-One Dehazing Network

Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, Dan Feng; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4770-4778

This paper proposes an image dehazing model built with a convolutional neural network (CNN), called All-in-One Dehazing Network (AOD-Net). It is designed based on a re-formulated atmospheric scattering model. Instead of estimating the transmission matrix and the atmospheric light separately as most previous models did, AOD-Net directly generates the clean image through a light-weight CNN. Such a novel end-to-end design makes it easy to embed AOD-Net into other deep models, e.g., Faster R-CNN, for improving high-level tasks on hazy images. Experimental results on both synthesized and natural hazy image datasets demonstrate our superior performance than the state-of-the-art in terms of PSNR, SSIM and the subjective visual quality. Furthermore, when concatenating AOD-Net with Faster R-CNN, we witness a large improvement of the object detection performance on hazy images.

****

Simultaneous Detection and Removal of High Altitude Clouds From an Image

Tushar Sandhan, Jin Young Choi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4779-4788

Interestingly, shape of the high-altitude clouds serves as a beacon for weather forecasting, so its detection is of vital importance. Besides these clouds often cause hindrance in an endeavor of satellites to inspect our world. Even thin clouds produce the undesired superposition of visual information, whose decomposition into the clear background and cloudy layer using a single satellite image is a highly ill-posed problem. In this work, we derive sophisticated image priors by thoroughly analyzing the properties of high-altitude clouds and geological images; and formulate a non-convex optimization scheme, which simultaneously detects and removes the clouds within a few seconds. Experimental results on real world RGB images demonstrate that the proposed method outperforms the other competitive methods by retaining the comprehensive background details and producing the precise shape of the cloudy layer.

****

Understanding Low- and High-Level Contributions to Fixation Prediction

Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, Matthias Bethge; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4789-4798

Understanding where people look in images is an important problem in computer vision. Despite significant research, it remains unclear to what extent human fixations can be predicted by low-level (contrast) compared to high-level (presence of objects) image features. Here we address this problem by introducing two novel models that use different feature spaces but the same readout architecture. Th e first model predicts human fixations based on deep neural network features trained on object recognition. This model sets a new state-of-the art in fixation prediction by achieving top performance in area under the curve metrics on the MIT300 hold-out benchmark (AUC = 88%, sAUC = 77%, NSS = 2.34). The second model uses purely low-level (isotropic contrast) features. This model achieves better performance than all models not using features pre-trained on object recognition, making it a strong baseline to assess the utility of high-level features. We then evaluate and visualize which fixations are better explained by low-level compared to high-level image features. Surprisingly we find that a substantial proportion of fixations are better explained by the simple low-level model than the state-of-the-art model. Comparing different features within the same powerful readout architecture allows us to better understand the relevance of low- versus high-level features in predicting fixation locations, while simultaneously achieving state-of-the-art saliency prediction.

****

Image Super-Resolution Using Dense Skip Connections

Tong Tong, Gen Li, Xiejie Liu, Qinquan Gao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4799-4807

Recent studies have shown that the performance of single-image super-resolution methods can be significantly boosted by using deep convolutional neural networks. In this study, we present a novel single-image super-resolution method by introducing dense skip connections in a very deep network. In the proposed network, the feature maps of each layer are propagated into all subsequent layers, providing an effective way to combine the low-level features and high-level features to boost the reconstruction performance. In addition, the dense skip connections in the network enable short paths to be built directly from the output to each layer, alleviating the vanishing-gradient problem of very deep networks. Moreover, deconvolution layers are integrated into the network to learn the upsampling filters and to speedup the reconstruction process. Further, the proposed method substantially reduces the number of parameters, enhancing the computational efficiency. We evaluate the proposed method using images from four benchmark datasets and set a new state of the art.

****

Convergence Analysis of MAP Based Blur Kernel Estimation

Sunghyun Cho, Seungyong Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4808-4816

One popular approach for blind deconvolution is to formulate a maximum a posteriori (MAP) problem with sparsity priors on the gradients of the latent image, and then alternatingly estimate the blur kernel and the latent image. While several successful MAP based methods have been proposed, there has been much controversy and confusion about their convergence, because sparsity priors have been shown to prefer blurry images to sharp natural images. In this paper, we revisit this problem and provide an analysis on the convergence of MAP based approaches. We first introduce a slight modification to a conventional joint energy function for blind deconvolution. The reformulated energy function yields the same alternating estimation process, but more clearly reveals how blind deconvolution works. We then show the energy function can actually favor the right solution instead of the no-blur solution under certain conditions, which explains the success of previous MAP based approaches. The reformulated energy function and our conditions for the convergence also provide a way to compare the qualities of different blur kernels, and we demonstrate its applicability to automatic blur kernel size selection, blur kernel estimation using light streaks, and defocus estimation.

****

Blob Reconstruction Using Unilateral Second Order Gaussian Kernels With Application to High-ISO Long-Exposure Image Denoising

Gang Wang, Carlos Lopez-Molina, Bernard De Baets; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4817-4825

Blob detection and image denoising are fundamental, and sometimes related, tasks in computer vision. In this paper, we propose a blob reconstruction method using scale-invariant normalized unilateral second order Gaussian kernels. Unlike ot

her blob detection methods, our method suppresses non-blob structures while also identifying blob parameters, i.e., position, prominence and scale, thereby facilitating blob reconstruction. We present an algorithm for high-ISO long-exposure noise removal that results from the combination of our blob reconstruction method and state-of-the-art denoising methods, i.e., the non-local means algorithm (NLM) and the color version of block-matching and 3-D filtering (CBM3D). Experiments on standard images corrupted by real high-ISO long-exposure noise and real-world noisy images demonstrate that our schemes incorporating the blob reduction procedure outperform both the original NLM and CBM3D.
****

Deep Generative Adversarial Compression Artifact Removal

Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, Alberto Del Bimbo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4826-4835

Compression artifacts arise in images whenever a lossy compression algorithm is applied. These artifacts eliminate details present in the original image, or add noise and small structures; because of these effects they make images less pleasant for the human eye, and may also lead to decreased performance of computer vision algorithms such as object detectors. To eliminate such artifacts, when decompressing an image, it is required to recover the original image from a disturbed version. To this end, we present a feed-forward fully convolutional residual network model trained using a generative adversarial framework. To provide a baseline, we show that our model can be also trained optimizing the Structural Similarity (SSIM), which is a better loss with respect to the simpler Mean Squared Error (MSE). Our GAN is able to produce images with more photorealistic details than MSE or SSIM based networks. Moreover we show that our approach can be used as a pre-processing step for object detection in case images are degraded by compression to a point that state-of-the art detectors fail. In this task, our GAN method obtains better performance than MSE or SSIM trained networks.
****

Online Multi-Object Tracking Using CNN-Based Single Object Tracker With Spatial-Temporal Attention Mechanism

Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, Nenghai Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4836-4845

In this paper, we propose a CNN-based framework for online MOT. This framework utilizes the merits of single object trackers in adapting appearance models and searching for target in the next frame. Simply applying single object tracker for MOT will encounter the problem in computational efficiency and drifted results caused by occlusion. Our framework achieves computational efficiency by sharing features and using ROI-Pooling to obtain individual features for each target. Some online learned target-specific CNN layers are used for adapting the appearance model for each target. In the framework, we introduce spatial-temporal attention mechanism (STAM) to handle the drift caused by occlusion and interaction among targets. The visibility map of the target is learned and used for inferring the spatial attention map. The spatial attention map is then applied to weight the features. Besides, the occlusion status can be estimated from the visibility map, which controls the online updating process via weighted loss on training samples with different occlusion statuses in different frames. It can be considered as temporal attention mechanism. The proposed algorithm achieves 34.3% and 46.0% in MOTA on challenging MOT15 and MOT16 benchmark dataset respectively.
****

Mutual Enhancement for Detection of Multiple Logos in Sports Videos

Yuan Liao, Xiaoqing Lu, Chengcui Zhang, Yongtao Wang, Zhi Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4846-4855

Detecting logo frequency and duration in sports videos provides sponsors an effective way to evaluate their advertising efforts. However, general-purposed object detection methods cannot address all the challenges in sports videos. In this paper, we propose a mutual-enhanced approach that can improve the detection of a logo through the information obtained from other simultaneously occurred logos.

In a Fast-RCNN-based framework, we first introduce a homogeneity-enhanced re-ranking method by analyzing the characteristics of homogeneous logos in each frame, including type repetition, color consistency, and mutual exclusion. Different from conventional enhance mechanism that improves the weak proposals with the dominant proposals, our mutual method can also enhance the relatively significant proposals with weak proposals. Mutual enhancement is also included in our frame propagation mechanism that improves logo detection by utilizing the continuity of logos across frames. We use a tennis video dataset and an associated logo collection for detection evaluation. Experiments show that the proposed method outperforms existing methods with a higher accuracy.

****

Referring Expression Generation and Comprehension via Attributes

Jingyu Liu, Liang Wang, Ming-Hsuan Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4856-4864

Referring expression is a kind of language expression that used for referring to particular objects.To make the expression without ambiguation, people often use attributes to describe the particular object. In this paper, we explore the role of attributes by incorporating them into both referring expression generation and comprehension. We first train an attribute learning model from visual objects and their paired descriptions. Then in the generation task, we take the learned attributes as the input into the generation model, thus the expressions are generated driven by both attributes and the previous words. For comprehension, we embed the learned attributes with visual features and semantics into the common space model, then the target object is retrieved based on its ranking distance in the common space. Experimental results on the three standard datasets, RefCOCO, RefCOCO+, and RefCOCOg show significant improvements over the baseline model, demonstrating that our methods are effective for both tasks.

****

RoomNet: End-To-End Room Layout Estimation

Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, Andrew Rabinovich; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4865-4874

This paper focuses on the task of room layout estimation from a monocular RGB image. Prior works break the problem into two sub-tasks: semantic segmentation of floor, walls, ceiling to produce layout hypotheses, followed by an iterative optimization step to rank these hypotheses. In contrast, we adopt a more direct formulation of this problem as one of estimating an ordered set of room layout keypoints. The room layout and the corresponding segmentation is completely specified given the locations of these ordered keypoints. We predict the locations of the room layout keypoints using RoomNet, an end-to-end trainable encoder-decoder network. On the challenging benchmark datasets Hedau and LSUN, we achieve state-of-the-art performance along with 200x to 600x speedup compared to the most recent work. Additionally, we present optional extensions to the RoomNet architecture such as including recurrent computations and memory units to refine the keypoint locations under the same parametric capacity.

****

SSH: Single Stage Headless Face Detector

Mahyar Najibi, Pouya Samangouei, Rama Chellappa, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4875-4884

We introduce the Single Stage Headless (SSH) face detector. Unlike two stage proposal-classification detectors, SSH detects faces in a single stage directly from the early convolutional layers in a classification network. SSH is headless. That is, it is able to achieve state-of-the-art results while removing the "head" of its underlying classification network -- i.e. all fully connected layers in the VGG-16 which contains a large number of parameters. Additionally, instead of relying on an image pyramid to detect faces with various scales, SSH is scale-invariant by design. We simultaneously detect faces with different scales in a single forward pass of the network, but from different layers. These properties make SSH fast and light-weight. Surprisingly, with a headless VGG-16, SSH beats the ResNet-101-based state-of-the-art on the WIDER dataset. Even though, unlike th e current state-of-the-art, SSH does not use an image pyramid and is 5X faster. Moreover, if an image pyramid is deployed, our light-weight network achieves state-of-the-art on all subsets of the WIDER dataset, improving the AP by 2.5%. SSH also reaches state-of-the-art results on the FDDB and Pascal-Faces datasets while using a small input size, leading to a speed of 50 frames/second on a GPU.

****

AnnArbor: Approximate Nearest Neighbors Using Arborescence Coding

Artem Babenko, Victor Lempitsky; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4885-4893

To compress large datasets of high-dimensional descriptors, modern quantization schemes learn multiple codebooks and then represent individual descriptors as combinations of codewords. Once the codebooks are learned, these schemes encode descriptors independently. In contrast to that, we present a new coding scheme that arranges dataset descriptors into a set of arborescence graphs, and then encodes non-root descriptors by quantizing their displacements with respect to their parent nodes. By optimizing the structure of arborescences, our coding scheme can decrease the quantization error considerably, while incurring only minimal overhead on the memory footprint and the speed of nearest neighbor search in the compressed dataset compared to the independent quantization. The advantage of the proposed scheme is demonstrated in a series of experiments with datasets of SIFT and deep descriptors.

****

Boosting Image Captioning With Attributes

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, Tao Mei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4894-4902

Automatically describing an image with a natural language has been an emerging challenge in both fields of computer vision and natural language processing. In this paper, we present Long Short-Term Memory with Attributes (LSTM-A) - a novel architecture that integrates attributes into the successful Convolutional Neural Networks (CNNs) plus Recurrent Neural Networks (RNNs) image captioning framework, by training them in an end-to-end manner. Particularly, the learning of attributes is strengthened by integrating inter-attribute correlations into Multiple Instance Learning (MIL). To incorporate attributes into captioning, we construct variants of architectures by feeding image representations and attributes into RNNs in different ways to explore the mutual but also fuzzy relationship between them. Extensive experiments are conducted on COCO image captioning dataset and our framework shows clear improvements when compared to state-of-the-art deep models. More remarkably, we obtain METEOR/CIDEr-D of 25.5%/100.2% on testing data of widely used and publicly available splits in [10] when extracting image representations by GoogleNet and achieve superior performance on COCO captioning Leaderboard.

****

Learning to Estimate 3D Hand Pose From Single RGB Images

Christian Zimmermann, Thomas Brox; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4903-4911

Low-cost consumer depth cameras and deep learning have enabled reasonable 3D hand pose estimation from single depth images. In this paper, we present an approach that estimates 3D hand pose from regular RGB images. This task has far more ambiguities due to the missing depth information. To this end, we propose a deep network that learns a network-implicit 3D articulation prior. Together with detected keypoints in the images, this network yields good estimates of the 3D pose. We introduce a large scale 3D hand pose dataset based on synthetic hand models for training the involved networks. Experiments on a variety of test sets, including one on sign language recognition, demonstrate the feasibility of 3D hand pose estimation on single color images.

****

Locally-Transferred Fisher Vectors for Texture Classification

Yang Song, Fan Zhang, Qing Li, Heng Huang, Lauren J. O'Donnell, Weidong Cai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4912-4920

Texture classification has been extensively studied in computer vision. Recent research shows that the combination of Fisher vector (FV) encoding and convolutional neural network (CNN) provides significant improvement in texture classification over the previous feature representation methods. However, by truncating the CNN model at the last convolutional layer, the CNN-based FV descriptors would not incorporate the full capability of neural networks in feature learning. In this study, we propose that we can further transform the CNN-based FV descriptors in a neural network model to obtain more discriminative feature representations. In particular, we design a locally-transferred Fisher vector (LFV) method, which involves a multi-layer neural network model containing locally connected layers to transform the input FV descriptors with filters of locally shared weights. The network is optimized based on the hinge loss of classification, and transferred FV descriptors are then used for image classification. Our results on three challenging texture image datasets show improved performance over the state of the art.

****

Object-Level Proposals

Jianxiang Ma, Anlong Ming, Zilong Huang, Xinggang Wang, Yu Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4921-4929

Edge and surface are two fundamental visual elements of an object. The majority of existing object proposal approaches utilize edge or edge-like cues to rank candidates, while we consider that the surface cue containing the 3D characteristic of objects should be captured effectively for proposals, which has been rarely discussed before. In this paper, an object-level proposal model is presented, which constructs an occlusion-based objectness taking the surface cue into account. Specifically, the better detection of occlusion edges is focused on to enrich the surface cue into proposals, namely, the occlusion-dominated fusion and normalization criterion are designed to obtain the approximately overall contour information, to enhance the occlusion edge map at utmost and thus boost proposals. Experimental results on the PASCAL VOC 2007 and MS COCO 2014 dataset demonstrate the effectiveness of our approach, which achieves around 6% improvement on the average recall than Edge Boxes at 1000 proposals and also leads to a modest gain on the performance of object detection.

****

Extreme Clicking for Efficient Object Annotation

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, Vittorio Ferrari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4930-4939

Manually annotating object bounding boxes is central to building computer vision datasets, and it is very time consuming (annotating ILSVRC [53] took 35s for one high-quality box [62]). It involves clicking on imaginary corners of a tight box around the object. This is difficult as these corners are often outside the actual object and several adjustments are required to obtain a tight box. We propose extreme clicking instead: we ask the annotator to click on four physical points on the object: the top, bottom, left- and right-most points. This task is more natural and these points are easy to find. We crowd-source extreme point annotations for PASCAL VOC 2007 and 2012 and show that (1) annotation time is only 7s per box, 5x faster than the traditional way of drawing boxes [62]; (2) the quality of the boxes is as good as the original ground-truth drawn the traditional way; (3) detectors trained on our annotations are as accurate as those trained on the original ground-truth. Moreover, our extreme clicking strategy not only yields box coordinates, but also four accurate boundary points. We show (4) how to incorporate them into GrabCut to obtain more accurate segmentations than those delivered when initializing it from bounding boxes; (5) semantic segmentations models trained on these segmentations outperform those trained on segmentations derived from bounding boxes.

****

WordSup: Exploiting Word Annotations for Character Based Text Detection

Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, Errui Ding; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4

940-4949

Imagery texts are usually organized as a hierarchy of several visual elements, i.e. characters, words, text lines and text blocks. Among these elements, character is the most basic one for various languages such as Western, Chinese, Japanese, mathematical expression and etc. It is natural and convenient to construct a common text detection engine based on character detectors. However, training character detectors requires a vast of location annotated characters, which are expensive to obtain. Actually, the existing real text datasets are mostly annotated in word or line level. To remedy this dilemma, we propose a weakly supervised framework that can utilize word annotations, either in tight quadrangles or the more loose bounding boxes, for character detector training. When applied in scene text detection, we are thus able to train a robust character detector by exploiting word annotations in the rich large-scale real scene text datasets, e.g. ICDAR15 [??] and COCO-text [??]. The character detector acts as a key role in the pipeline of our text detection engine. It achieves the state-of-the-art performance on several challenging scene text detection benchmarks. We also demonstrate the flexibility of our pipeline by various scenarios, including deformed text detection and math expression recognition.

****

Illuminating Pedestrians via Simultaneous Detection & Segmentation

Garrick Brazil, Xi Yin, Xiaoming Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4950-4959

Pedestrian detection is a critical problem in computer vision with significant impact on safety in urban autonomous driving. In this work, we explore how semantic segmentation can be used to boost pedestrian detection accuracy while having little to no impact on network efficiency. We propose a segmentation infusion network to enable joint supervision on semantic segmentation and pedestrian detection. When placed properly, the additional supervision helps guide features in shared layers to become more sophisticated and helpful for the downstream pedestrian detector. Using this approach, we find weakly annotated boxes to be sufficient for considerable performance gains. We provide an in-depth analysis to demonstrate how shared layers are shaped by the segmentation supervision. In doing so, we show that the resulting feature maps become more semantically meaningful and robust to shape and occlusion. Overall, our simultaneous detection and segmentation framework achieves a considerable gain over the state-of-the-art on the Caltech pedestrian dataset, competitive performance on KITTI, and executes 2x faster than competitive methods.

****

Generalized Orderless Pooling Performs Implicit Salient Matching

Marcel Simon, Yang Gao, Trevor Darrell, Joachim Denzler, Erik Rodner; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4960-4969

Most recent CNN architectures use average pooling as a final feature encoding step. In the field of fine-grained recognition, however, recent global representations like bilinear pooling offer improved performance. In this paper, we generalize average and bilinear pooling to "alpha-pooling", allowing for learning the pooling strategy during training. In addition, we present a novel way to visualize decisions made by these approaches. We identify parts of training images having the highest influence on the prediction of a given test image. This allows for justifying decisions to users and also for analyzing the influence of semantic parts. For example, we can show that the higher capacity VGG16 model focuses much more on the bird's head than, e.g., the lower-capacity VGG-M model when recognizing fine-grained bird categories. Both contributions allow us to analyze the difference when moving between average and bilinear pooling. In addition, experiments show that our generalized approach can outperform both across a variety of standard datasets.

****

Exploiting Spatial Structure for Localizing Manipulated Image Regions

Jawadul H. Bappy, Amit K. Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, B. S. Manjunath; Proceedings of the IEEE International Conference on Computer Vision (I

CCV), 2017, pp. 4970-4979

The advent of high-tech journaling tools facilitates an image to be manipulated in a way that can easily evade state-of-the-art image tampering detection approaches. The recent success of the deep learning approaches in different recognition tasks inspires us to develop a high confidence detection framework which can localize manipulated regions in an image. Unlike semantic object segmentation where all meaningful regions (objects) are segmented, the localization of image manipulation focuses only the possible tampered region which makes the problem even more challenging. In order to formulate the framework, we employ a hybrid CNN-LSTM model to capture discriminative features between manipulated and non-manipulated regions. One of the key properties of manipulated regions is that they exhibit discriminative features in boundaries shared with neighboring non-manipulated pixels. Our motivation is to learn the boundary discrepancy, i.e., the spatial structure, between manipulated and non-manipulated regions with the combination of LSTM and convolution layers. We perform end-to-end training of the network to learn the parameters through back-propagation given groundtruth mask information. The overall framework is capable of detecting different types of image manipulations, including copy-move, removal and splicing. Our model shows promising results in localizing manipulated regions, which is demonstrated through rigorous experimentation on three diverse datasets.

****

RDFNet: RGB-D Multi-Level Residual Feature Fusion for Indoor Semantic Segmentation

Seong-Jin Park, Ki-Sang Hong, Seungyong Lee; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4980-4989

In multi-class indoor semantic segmentation using RGB-D data, it has been shown that incorporating depth feature into RGB feature is helpful to improve segmentation accuracy. However, previous studies have not fully exploited the potentials of multi-modal feature fusion, e.g., simply concatenating RGB and depth features or averaging RGB and depth score maps. To learn the optimal fusion of multi-modal features, this paper presents a novel network that extends the core idea of residual learning to RGB-D semantic segmentation. Our network effectively captures multi-level RGB-D CNN features by including multi-modal feature fusion blocks and multi-level feature refinement blocks. Feature fusion blocks learn residual RGB and depth features and their combinations to fully exploit the complementary characteristics of RGB and depth data. Feature refinement blocks learn the combination of fused features from multiple levels to enable high-resolution prediction. Our network can efficiently train discriminative multi-level features from each modality end-to-end by taking full advantage of skip-connections. Our comprehensive experiments demonstrate that the proposed architecture achieves the state-of-the-art accuracy on two challenging RGB-D indoor datasets, NYUDv2 and SUN RGB-D.

****

The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, Peter Kontschieder; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4990-4999

The Mapillary Vistas Dataset is a novel, large-scale street-level image dataset containing 25,000 high-resolution images annotated into 66 object categories with additional, instance-specific labels for 37 classes. Annotation is performed in a dense and fine-grained style by using polygons for delineating individual objects. Our dataset is 5x larger than the total amount of fine annotations for Cityscapes and contains images from all around the world, captured at various conditions regarding weather, season and daytime. Images come from different imaging devices (mobile phones, tablets, action cameras, professional capturing rigs) and differently experienced photographers. In such a way, our dataset has been designed and compiled to cover diversity, richness of detail and geographic extent. As default benchmark tasks, we define semantic image segmentation and instance-specific image segmentation, aiming to significantly further the development of state-of-the-art methods for visual road-scene understanding.

****

Self-Organized Text Detection With Minimal Post-Processing via Border Learning

Yue Wu, Prem Natarajan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5000-5009

In this paper we propose a new solution to the text detection problem via border learning. Specifically, we make four major contributions: 1) We analyze the insufficiencies of the classic non-text and text settings for text detection. 2) We introduce the border class to the text detection problem for the first time, and validate that the decoding process is largely simplified with the help of text border. 3) We collect and release a new text detection ppt dataset containing 10,692 images with non-text, border, and text annotations. 4) We develop a lightweight (only 0.28M parameters), fully convolutional network (FCN) to effectively learn borders in text images. The results of our extensive experiments show that the proposed solution achieves comparable performance, and often outperforms state-of-the-art approaches on standard benchmarks--even though our solution only requires minimal post-processing to parse a bounding box from a detected text map, while others often require heavy post-processing.

****

Sparse Exact PGA on Riemannian Manifolds

Monami Banerjee, Rudrasis Chakraborty, Baba C. Vemuri; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5010-5018

Principal Component Analysis (PCA) is a widely popular dimensionality reduction technique for vector-valued inputs. In the past decade, a nonlinear generalization of PCA, called the Principal Geodesic Analysis (PGA) was developed to tackle data that lie on a smooth manifold. PGA suffers from the same problem as PCA in that, in both the methods, each Principal Component (PC) is a linear combination of the original variables. This makes it very difficult to interpret the PCs especially in high dimensions. This lead to the introduction of sparse PCA (SPCA) in the vector space input case. In this paper, we present a novel generalization of SPCA, called sparse exact PGA (SEPGA) that can cope with manifold-valued input data and respect the intrinsic geometry of the underlying manifold. Sparsity has the advantage of not only easy interpretability but also computational efficiency. We achieve this by formulating the PGA problem as a minimization of the projection error in conjunction with sparsity constraints enforced on the principal vectors post isomorphic mapping to Rm, where m is the dimension of the manifold on which the data reside. Further, for constant curvature smooth manifolds, we use analytic formulae for the projection error leading to an efficient solution to the SEPGA problem. We present extensive experimental results demonstrating the performance of SEPGA to achieve very good sparse principal components without sacrificing the accuracy of reconstruction. This makes the representation of manifold-valued data using SEPGA accurate and efficient.

****

Tensor RPCA by Bayesian CP Factorization With Complex Noise

Qiong Luo, Zhi Han, Xi'ai Chen, Yao Wang, Deyu Meng, Dong Liang, Yandong Tang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5019-5028

The RPCA model has achieved good performances in various applications. However, two defects limit its effectiveness. Firstly, it is designed for dealing with data in matrix form, which fails to exploit the structure information of higher order tensor data in some pratical situations. Secondly, it adopts L1-norm to tackle noise part which makes it only valid for sparse noise. In this paper, we propose a tensor RPCA model based on CP decomposition and model data noise by Mixture of Gaussians (MoG). The use of tensor structure to raw data allows us to make full use of the inherent structure priors, and MoG is a general approximator to any blends of consecutive distributions, which makes our approach capable of regaining the low dimensional linear subspace from a wide range of noises or their mixture. The model is solved by a new proposed algorithm inferred under a variational Bayesian framework. The superiority of our approach over the existing state-of-the-art approaches is demonstrated by extensive experiments on both of synthetic and real data.

****

Multimodal Gaussian Process Latent Variable Models With Harmonization

Guoli Song, Shuhui Wang, Qingming Huang, Qi Tian; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5029-5037

In this work, we address multimodal learning problem with Gaussian process latent variable models (GPLVMs) and their application to cross-modal retrieval. Existing GPLVM based studies generally impose individual priors over the model parameters and ignore the intrinsic relations among these parameters. Considering the strong complementarity between modalities, we propose a novel joint prior over the parameters for multimodal GPLVMs to propagate multimodal information in both kernel hyperparameter spaces and latent space. The joint prior is formulated as a harmonization constraint on the model parameters, which enforces the agreement among the modality-specific GP kernels and the similarity in the latent space. We incorporate the harmonization mechanism into the learning process of multimodal GPLVMs. The proposed methods are evaluated on three widely used multimodal datasets for cross-modal retrieval. Experimental results show that the harmonization mechanism is beneficial to the GPLVM algorithms for learning non-linear correlation among heterogeneous modalities.

****

Segmentation-Aware Convolutional Networks Using Local Attention Masks

Adam W. Harley, Konstantinos G. Derpanis, Iasonas Kokkinos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5038-5047

We introduce an approach to integrate segmentation information within a convolutional neural network (CNN). This counter-acts the tendency of CNNs to smooth information across regions and increases their spatial precision. To obtain segmentation information, we set up a CNN to provide an embedding space where region co-membership can be estimated based on Euclidean distance. We use these embeddings to compute a local attention mask relative to every neuron position. We incorporate such masks in CNNs and replace the convolution operation with a "segmentation-aware" variant that allows a neuron to selectively attend to inputs coming from its own region. We call the resulting network a segmentation-aware CNN because it adapts its filters at each image point according to local segmentation cues, while at the same time remaining fully-convolutional. We demonstrate the merit of our method on two widely different dense prediction tasks, that involve classification (semantic segmentation) and regression (optical flow). Our results show that in semantic segmentation we can replace DenseCRF inference with a cascade of segmentation-aware filters, and in optical flow we obtain clearly sharper responses than the ones obtained with comparable networks that do not use segmentation. In both cases segmentation-aware convolution yields systematic improvements over strong baselines.

****

Rotation Equivariant Vector Field Networks

Diego Marcos, Michele Volpi, Nikos Komodakis, Devis Tuia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5048-5057

In many computer vision tasks, we expect a particular behavior of the output with respect to rotations of the input image. If this relationship is explicitly encoded, instead of treated as any other variation, the complexity of the problem is decreased, leading to a reduction in the size of the required model. We propose Rotation Equivariant vector field Networks (RotEqNet) to encode rotation equivariance and invariance into Convolutional Neural Networks (CNNs). Each convolutional filter is applied at multiple orientations and returns a vector field that represents the magnitude and angle of the highest scoring orientation at every spatial location. A modified convolution operator using vector fields as inputs and filters can then be applied to obtain deep architectures. We test RotEqNet on several problems requiring different responses with respect to the inputs' rotation: image classification, biomedical image segmentation, orientation estimation and patch matching. In all cases, we show that RotEqNet offers very compact models in terms of number of parameters and provides results in line to those of networks orders of magnitude larger.

****

ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression
Jian-Hao Luo, Jianxin Wu, Weiyao Lin; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5058-5066
We propose an efficient and unified framework, namely ThiNet, to simultaneously accelerate and compress CNN models in both training and inference stages. We focus on the filter level pruning, i.e., the whole filter would be discarded if it is less important. Our method does not change the original network structure, thus it can be perfectly supported by any off-the-shelf deep learning libraries. We formally establish filter pruning as an optimization problem, and reveal that we need to prune filters based on statistics information computed from its next layer, not the current layer, which differentiates ThiNet from existing methods. Experimental results demonstrate the effectiveness of this strategy, which has advanced the state-of-the-art. We also show the performance of ThiNet on ILSVRC-12 benchmark. ThiNet achieves 3.31x FLOPs reduction and 16.63x compression on VGG-16, with only 0.52% top-5 accuracy drop. Similar experiments with ResNet-50 reveal that even for a compact network, ThiNet can also reduce more than half of the parameters and FLOPs, at the cost of roughly 1% top-5 accuracy drop. Moreover, the original VGG-16 model can be further pruned into a very small model with only 5.05MB model size, preserving AlexNet level accuracy but showing much stronger generalization ability.
****
AutoDIAL: Automatic DomaIn Alignment Layers
Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, Samuel Rota Bulo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5067-5075
Classifiers trained on given databases perform poorly when tested on data acquired in different settings. This is explained in domain adaptation through a shift among distributions of the source and target domains. Attempts to align them have traditionally resulted in works reducing the domain shift by introducing appropriate loss terms, measuring the discrepancies between source and target distributions, in the objective function. Here we take a different route, proposing to align the learned representations by embedding in any given network specific Domain Alignment Layers, designed to match the source and target feature distributions to a reference one. Opposite to previous works which define a priori in which layers adaptation should be performed, our method is able to automatically learn the degree of feature alignment required at different levels of the deep network. Thorough experiments on different public benchmarks, in the unsupervised setting, confirm the power of our approach.
****
Focusing Attention: Towards Accurate Text Recognition in Natural Images
Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, Shuigeng Zhou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5076-5084
Scene text recognition has been a hot research topic in computer vision due to its various applications. The state of the art is the attention-based encoder-decoder framework that learns the mapping between input images and output sequences in a purely data-driven way. However, we observe that existing attention-based methods perform poorly on complicated and/or low-quality images. One major reason is that existing methods cannot get accurate alignments between feature areas and targets for such images. We call this phenomenon "attention drift". To tackle this problem, in this paper we propose the FAN (the abbreviation of Focusing Attention Network) method that employs a focusing attention mechanism to automatically draw back the drifted attention. FAN consists of two major components: an attention network (AN) that is responsible for recognizing character targets as in the existing methods, and a focusing network (FN) that is responsible for adjusting attention by evaluating whether AN pays attention properly on the target areas in the images. Furthermore, different from the existing methods, we adopt a ResNet-based network to enrich deep representations of scene text images. Extensive experiments on various benchmarks, including the IIIT5k, SVT and ICDAR datasets, show that the FAN method substantially outperforms the existing methods.

****

Unsupervised Object Segmentation in Video by Efficient Selection of Highly Probable Positive Features

Emanuela Haller, Marius Leordeanu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5085-5093

We address an essential problem in computer vision, that of unsupervised foreground object segmentation in video, where a main object of interest in a video sequence should be automatically separated from its background. An efficient solution to this task would enable large-scale video interpretation at a high semantic level in the absence of the costly manual labeling. We propose an efficient unsupervised method for generating foreground object soft masks based on automatic selection and learning from highly probable positive features. We show that such features can be selected efficiently by taking into consideration the spatio-temporal appearance and motion consistency of the object in the video sequence. We also emphasize the role of the contrasting properties between the foreground object and its background. Our model is created over several stages: we start from pixel level analysis and move to descriptors that consider information over groups of pixels combined with efficient motion analysis. We also prove theoretical properties of our unsupervised learning method, which under some mild constraints is guaranteed to learn the correct classifier even in the unsupervised case. We achieve competitive and even state of the art results on the challenging Youtube-Objects and SegTrack datasets, while being at least one order of magnitude faster than the competition. We believe that the strong performance of our method, along with its theoretical properties, constitute a solid step towards solving unsupervised discovery in video.

****

Nonparametric Variational Auto-Encoders for Hierarchical Representation Learning

Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, Eric P. Xing; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5094-5102

The recently developed variational autoencoders (VAEs) have proved to be an effective confluence of the rich representational power of neural networks with Bayesian methods. However, most work on VAEs use a rather simple prior over the latent variables such as standard normal distribution, thereby restricting its applications to relatively simple phenomena. In this work, we propose hierarchical nonparametric variational autoencoders, which combines tree-structured Bayesian nonparametric priors with VAEs, to enable infinite flexibility of the latent representation space. Both the neural parameters and Bayesian priors are learned jointly using tailored variational inference. The resulting model induces a hierarchical structure of latent semantic concepts underlying the data corpus, and infers accurate representations of data instances. We apply our model in video representation learning. Our method is able to discover highly interpretable activity hierarchies, and obtain improved clustering accuracy and generalization capacity based on the learned rich representations.

****

Dense and Low-Rank Gaussian CRFs Using Deep Embeddings

Siddhartha Chandra, Nicolas Usunier, Iasonas Kokkinos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5103-5112

In this work we introduce a structured prediction model that endows the Deep Gaussian Conditional Random Field (G-CRF) with a densely connected graph structure. We keep memory and computational complexity under control by expressing the pairwise interactions as inner products of low-dimensional, learnable embeddings. The G-CRF system matrix is therefore low-rank, allowing us to solve the resulting system in a few milliseconds on the GPU by using conjugate gradients. As in G-CRF, inference is exact, the unary and pairwise terms are jointly trained end-to-end by using analytic expressions for the gradients, while we also develop even faster, Potts-type variants of our embeddings. We show that the learned embeddings capture pixel-to-pixel affinities in a task-specific manner, while our approach achieves state of the art results on three challenging benchmarks, namely semantic segmentation, human part segmentation, and saliency estimation. Our implem entation is fully GPU based, built on top of the Caffe library, and is available at https://github.com/siddharthachandra/gcrf-v2.0

****

A Multimodal Deep Regression Bayesian Network for Affective Video Content Analyses

Quan Gan, Shangfei Wang, Longfei Hao, Qiang Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5113-5122

The inherent dependencies between visual elements and aural elements are crucial for affective video content analyses, yet have not been successfully exploited. Therefore, we propose a multimodal deep regression Bayesian network (MMDRBN) to capture the dependencies between visual elements and aural elements for affective video content analyses. The regression Bayesian network (RBN) is a directed graphical model consisting of one latent layer and one visible layer. Due to the explaining away effect in Bayesian networks (BN), RBN is able to capture both the dependencies among the latent variables given the observation and the dependencies among visible variables. We propose a fast learning algorithm to learn the RBN. For the MMDRBN, first, we learn several RBNs layer-wisely from visual modality and audio modality respectively. Then we stack these RBNs and obtain two deep networks. After that, a joint representation is extracted from the top layers of the two deep networks, and thus captures the high order dependencies between visual modality and audio modality. In order to predict the valence or arousal score of video contents, we initialize a feed-forward inference network from the MMDRBN whose inference is intractable by minimizing the KullbackLeibler (KL)divergence between the two networks. The back propagation algorithm is adopted for finetuning the inference network. Experimental results on the LIRIS-ACCEDE database demonstrate that the proposed MMDRBN successfully captures the dependencies between visual and audio elements, and thus achieves better performance compared with state of the art work.

****

Moving Object Detection in Time-Lapse or Motion Trigger Image Sequences Using Low-Rank and Invariant Sparse Decomposition

Moein Shakeri, Hong Zhang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5123-5131

Low-rank and sparse representation based methods have attracted wide attention in background subtraction and moving object detection, where moving objects in the scene are modeled as pixel-wise sparse outliers. Since in real scenarios moving objects are also structurally sparse, recently researchers have attempted to extract moving objects using structured sparse outliers. Although existing methods with structured sparsity-inducing norms produce promising results, they are still vulnerable to various illumination changes that frequently occur in real environments, specifically for time-lapse image sequences where assumptions about sparsity between images such as group sparsity are not valid. In this paper, we first introduce a prior map obtained by illumination invariant representation of images. Next, we propose a low-rank and invariant sparse decomposition using the prior map to detect moving objects under significant illumination changes. Experiments on challenging benchmark datasets demonstrate the superior performance of our proposed method under complex illumination changes.

****

A Multilayer-Based Framework for Online Background Subtraction With Freely Moving Cameras

Yizhe Zhu, Ahmed Elgammal; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5132-5141

The exponentially increasing use of moving platforms for video capture introduces the urgent need to develop the general background subtraction algorithms with the capability to deal with the moving background. In this paper, we propose a multilayer-based framework for online background subtraction for videos captured by moving cameras. Unlike the previous treatments of the problem, the proposed method is not restricted to binary segmentation of background and foreground, but formulates it as a multi-label segmentation problem by modeling multiple foreground objects in different layers when they appear simultaneously in the scene. W e assign an independent processing layer to each foreground object, as well as the background, where both motion and appearance models are estimated, and a probability map is inferred using a Bayesian filtering framework. Finally, Multi-label Graph-cut on Markov Random Field is employed to perform pixel-wise labeling. Extensive evaluation results show that the proposed method outperforms state-of-the-art methods on challenging video sequences.

****

Dynamic Label Graph Matching for Unsupervised Video Re-Identification

Mang Ye, Andy J. Ma, Liang Zheng, Jiawei Li, Pong C. Yuen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5142-5150

Label estimation is an important component in an unsupervised person re-identification (re-ID) system. This paper focuses on cross-camera label estimation, which can be subsequently used in feature learning to learn robust re-ID models. Specifically, we propose to construct a graph for samples in each camera, and then graph matching scheme is introduced for cross-camera labeling association. While labels directly output from existing graph matching methods may be noisy and inaccurate due to significant cross-camera variations, this paper propose a dynamic graph matching (DGM) method. DGM iteratively updates the image graph and the label estimation process by learning a better feature space with intermediate estimated labels. DGM is advantageous in two aspects: 1) the accuracy of estimated labels is improved significantly with the iterations; 2) DGM is robust to noisy initial training data. Extensive experiments conducted on three benchmarks including the large-scale MARS dataset show that DGM yields competitive performance to fully supervised baselines, and outperforms competing unsupervised learning methods.

****

Spatiotemporal Modeling for Crowd Counting in Videos

Feng Xiong, Xingjian Shi, Dit-Yan Yeung; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5151-5159

Region of Interest (ROI) crowd counting can be formulated as a regression problem of learning a mapping from an image or a video frame to a crowd density map. Recently, convolutional neural network (CNN) models have achieved promising results for crowd counting. However, even when dealing with video data, CNN-based methods still consider each video frame independently, ignoring the strong temporal correlation between neighboring frames. To exploit the otherwise very useful temporal information in video sequences, we propose a variant of a recent deep learning model called convolutional LSTM (ConvLSTM) for crowd counting. Unlike the previous CNN-based methods, our method fully captures both spatial and temporal dependencies. Furthermore, we extend the ConvLSTM model to a bidirectional ConvLSTM model which can access long-range information in both directions. Extensive experiments using four publicly available datasets demonstrate the reliability of our approach and the effectiveness of incorporating temporal information to boost the accuracy of crowd counting. In addition, we also conduct some transfer learning experiments to show that once our model is trained on one dataset, its learning experience can be transferred easily to a new dataset which consists of only very few video frames for model adaptation.

****

Personalized Cinemagraphs Using Semantic Understanding and Collaborative Learning

Tae-Hyun Oh, Kyungdon Joo, Neel Joshi, Baoyuan Wang, In So Kweon, Sing Bing Kang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5160-5169

Cinemagraphs are a compelling way to convey dynamic aspects of a scene. In these media, dynamic and still elements are juxtaposed to create an artistic and narrative experience. Creating a high-quality, aesthetically pleasing cinemagraph requires isolating objects in a semantically meaningful way and then selecting good start times and looping periods for those objects to minimize visual artifacts (such a tearing). To achieve this, we present a new technique that uses object recognition and semantic segmentation as part of an optimization method to automatically create cinemagraphs from videos that are both visually appealing and se mantically meaningful. Given a scene with multiple objects, there are many cinemagraphs one could create. Our method evaluates these multiple candidates and presents the best one, as determined by a model trained to predict human preferences in a collaborative way. We demonstrate the effectiveness of our approach with multiple results and a user study.

****

What Is Around the Camera?

Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5170-5178

How much does a single image reveal about the environment it was taken in? In this paper, we investigate how much of that information can be retrieved from a foreground object, combined with the background (i.e. the visible part of the environment). Assuming it is not perfectly diffuse, the foreground object acts as a complexly shaped and far-from-perfect mirror. An additional challenge is that its appearance confounds the light coming from the environment with the unknown materials it is made of. We propose a learning-based approach to predict the environment from multiple reflectance maps that are computed from approximate surface normals. The proposed method allows us to jointly model the statistics of environments and material properties. We train our system from synthesized training data, but demonstrate its applicability to real-world data. Interestingly, our analysis shows that the information obtained from objects made out of multiple materials often is complementary and leads to better performance.

****

Weakly-Supervised Learning of Visual Relations

Julia Peyre, Josef Sivic, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5179-5188

This paper introduces a novel approach for modeling visual relations between pairs of objects. We call relation a triplet of the form (subject, predicate, object) where the predicate is typically a preposition (eg. 'under', 'in front of') or a verb ('hold', 'ride') that links a pair of objects (subject, object). Learning such relations is challenging as the objects have different spatial configurations and appearances depending on the relation in which they occur. Another major challenge comes from the difficulty to get annotations, especially at box-level, for all possible triplets, which makes both learning and evaluation difficult. The contributions of this paper are threefold. First, we design strong yet flexible visual features that encode the appearance and spatial configuration for pairs of objects. Second, we propose a weakly-supervised discriminative clustering model to learn relations from image-level labels only. Third we introduce a new challenging dataset of unusual relations (UnRel) together with an exhaustive annotation, that enables accurate evaluation of visual relation retrieval. We show experimentally that our model results in state-of-the-art results on the visual relationship dataset significantly improving performance on previously unseen relations (zero-shot learning), and confirm this observation on our newly introduced UnRel dataset.

****

BIER - Boosting Independent Embeddings Robustly

Michael Opitz, Georg Waltner, Horst Possegger, Horst Bischof; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5189-5198

Learning similarity functions between image pairs with deep neural networks yields highly correlated activations of large embeddings. In this work, we show how to improve the robustness of embeddings by exploiting independence in ensembles. We divide the last embedding layer of a deep network into an embedding ensemble and formulate training this ensemble as an online gradient boosting problem. Each learner receives a reweighted training sample from the previous learners. This leverages large embedding sizes more effectively by significantly reducing correlation of the embedding and consequently increases retrieval accuracy of the embedding. Our method does not introduce any additional parameters and works with any differentiable loss function. We evaluate our metric learning method on image retrieval tasks and show that it improves over state-of-the-art methods on th e CUB-200-2011, Cars-196, Stanford Online Products, In-Shop Clothes Retrieval and VehicleID datasets by a significant margin.

****

3D Graph Neural Networks for RGBD Semantic Segmentation

Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5199-5208

RGBD semantic segmentation requires joint reasoning about 2D appearance and 3D geometric information. In this paper we propose a 3D graph neural network (3DGNN) that builds a k-nearest neighbor graph on top of 3D point cloud. Each node in the graph corresponds to a set of points and is associated with a hidden representation vector initialized with an appearance feature extracted by a unary CNN from 2D images. Relying on recurrent functions, every node dynamically updates its hidden representation based on the current status and incoming messages from its neighbors. This propagation model is unrolled for a certain number of time steps and the final per-node representation is used for predicting the semantic class of each pixel. We use back-propagation through time to train the model. Extensive experiments on NYUD2 and SUN-RGBD datasets demonstrate the effectiveness of our approach.

****

Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition

Heliang Zheng, Jianlong Fu, Tao Mei, Jiebo Luo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5209-5217

Recognizing fine-grained categories (e.g., bird species) highly relies on discriminative part localization and part-based fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that part localization (e.g., head of a bird) and fine-grained feature learning (e.g., head shape) are mutually correlated. In this paper, we propose a novel part learning approach by a multi-attention convolutional neural network (MA-CNN), where part generation and feature learning can reinforce each other. MA-CNN consists of convolution, channel grouping and part classification sub-networks. The channel grouping network takes as input feature channels from convolutional layers, and generates multiple parts by clustering, weighting and pooling from spatially-correlated channels. The part classification network further classifies an image by each individual part, through which more discriminative fine-grained features can be learned. Two losses are proposed to guide the multi-task learning of channel grouping and part classification, which encourages MA-CNN to generate more discriminative parts from feature channels and learn better fine-grained features from parts in a mutual reinforced way. MA-CNN does not need bounding box/part annotation and can be trained end-to-end. We incorporate the learned parts from MA-CNN with part-CNN for recognition, and show the best performances on three challenging published fine-grained datasets, e.g., CUB-Birds, FGVC-Aircraft and Stanford-Cars.

****

Learning 3D Object Categories by Looking Around Them

David Novotny, Diane Larlus, Andrea Vedaldi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5218-5227

Traditional approaches for learning 3D object categories use either synthetic data or manual supervision. In this paper, we propose a method which does not require manual annotations and is instead cued by observing objects from a moving vantage point. Our system builds on two innovations: a Siamese viewpoint factorization network that robustly aligns different videos together without explicitly comparing 3D shapes; and a 3D shape completion network that can extract the full shape of an object from partial observations. We also demonstrate the benefits of configuring networks to perform probabilistic predictions as well as of geometry-aware data augmentation schemes. We obtain state-of-the-art results on publicly-available benchmarks.

****

Quantitative Evaluation of Confidence Measures in a Machine Learning World

Matteo Poggi, Fabio Tosi, Stefano Mattoccia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5228-5237

Confidence measures aim at detecting unreliable depth measurements and play an important role for many purposes and in particular, as recently shown, to improve stereo accuracy. This topic has been thoroughly investigated by Hu and Mordohai in 2010 (and 2012) considering 17 confidence measures and two local algorithms on the two datasets available at that time. However, since then major breakthroughs happened in this field: the availability of much larger and challenging datasets, novel and more effective stereo algorithms including ones based on deep-learning and confidence measures leveraging on machine learning techniques. Therefore, this paper aims at providing an exhaustive and updated review and quantitative evaluation of 52 (actually, 76 considering variants) state-of-the-art confidence measures - focusing on recent ones mostly based on random-forests and deep-learning - with three algorithms on the challenging datasets available today. Moreover we deal with problems inherently induced by learning-based confidence measures. How are these methods able to generalize to new data? How a specific training improves their effectiveness? How more effective confidence measures can actually improve the overall stereo accuracy?

****

Towards End-To-End Text Spotting With Convolutional Recurrent Neural Networks

Hui Li, Peng Wang, Chunhua Shen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5238-5246

In this work, we jointly address the problem of text detection and recognition in natural scene images based on convolutional recurrent neural networks. We propose a unified network that simultaneously localizes and recognizes text with a single forward pass, avoiding intermediate processes, such as image cropping, feature re-calculation, word separation, and character grouping. In contrast to existing approaches that consider text detection and recognition as two distinct tasks and tackle them one by one, the proposed framework settles these two tasks concurrently. The whole framework can be trained end-to-end, requiring only images, ground-truth bounding boxes and text labels. The convolutional features are calculated only once and shared by both detection and recognition, which saves processing time. Through multi-task training, the learned features become more informative and improves the overall performance. Our proposed method has achieved competitive performance on several benchmark datasets.

****

DeepSetNet: Predicting Sets With Deep Neural Networks

S. Hamid Rezatofighi, Vijay Kumar B G, Anton Milan, Ehsan Abbasnejad, Anthony Dick, Ian Reid; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5247-5256

This paper addresses the task of set prediction using deep learning. This is important because the output of many computer vision tasks, including image tagging and object detection, are naturally expressed as sets of entities rather than vectors. As opposed to a vector, the size of a set is not fixed in advance, and it is invariant to the ordering of entities within it. We define a likelihood for a set distribution and learn its parameters using a deep neural network. We also derive a loss for predicting a discrete distribution corresponding to set cardinality. Set prediction is demonstrated on the problem of multi-class image classification. Moreover, we show that the proposed cardinality loss can also trivially be applied to the tasks of object counting and pedestrian detection. Our approach outperforms existing methods in all three cases on standard datasets.

****

Learning From Video and Text via Large-Scale Discriminative Clustering

Antoine Miech, Jean-Baptiste Alayrac, Piotr Bojanowski, Ivan Laptev, Josef Sivic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5257-5266

Discriminative clustering has been successfully applied to a number of weakly supervised learning tasks. Such applications include person and action recognition, text-to-video alignment, object co-segmentation and colocalization in videos and images. One drawback of discriminative clustering, however, is its limited sc alability. We address this issue and propose an online optimization algorithm based on the Block-Coordinate Frank-Wolfe algorithm. We apply the proposed method to the problem of weakly supervised learning of actions and actors from movies together with corresponding movie scripts. The scaling up of the learning problem to 66 feature-length movies enables us to significantly improve weakly supervised action recognition.

****

TALL: Temporal Activity Localization via Language Query

Jiyang Gao, Chen Sun, Zhenheng Yang, Ram Nevatia; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5267-5275

This paper focuses on temporal localization of actions from untrimmed videos. Existing methods typically involve training classifiers for a pre-defined list of actions and applying the classifiers in a sliding window fashion. However, activities in the wild consist of a wide combination of actors, actions and objects; it is difficult to design a proper activity list that meets users' needs. We propose to localize activities by natural language queries. Temporal Activity Localization via Language (TALL) is challenging as it requires: (1) suitable design of text and video representations to allow cross-modal matching of actions and language queries; (2) ability to locate actions accurately given features from sliding windows of limited granularity. We propose a novel Cross-modal Temporal Regression Localizer (CTRL) to jointly model text query and video clips, output alignment scores and location regression results for candidate clips. For evaluation, we adopt TaCoS dataset, and build a new dataset for this task on top of Charades by adding sentence temporal annotations, called Charades-STA. Experimental results show that CTRL outperforms previous methods significantly on both datasets.

****

End-To-End Face Detection and Cast Grouping in Movies Using Erdos-Renyi Clustering

SouYoung Jin, Hang Su, Chris Stauffer, Erik Learned-Miller; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5276-5285

We present an end-to-end system for detecting and clustering faces by identity in full-length movies. Unlike works that start with a predefined set of detected faces, we consider the end-to-end problem of detection and clustering together. We make three separate contributions. First, we combine a state-of-the-art face detector with a generic tracker to extract high quality face tracklets. We then introduce a novel clustering method, motivated by the classic graph theory results of Erdos and Renyi. It is based on the observations that large clusters can be fully connected by joining just a small fraction of their point pairs, while just a single connection between two different people can lead to poor clustering results. This suggests clustering using a verification system with very few false positives but perhaps moderate recall. We introduce a novel verification method, rank-1 counts verification, that has this property, and use it in a link-based clustering scheme. Finally, we define a novel end-to-end detection and clustering evaluation metric allowing us to assess the accuracy of the entire end-to-end system. We present state-of-the-art results on multiple video data sets and also on standard face databases.

****

Active Decision Boundary Annotation With Deep Generative Models

Miriam Huijser, Jan C. van Gemert; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5286-5295

This paper is on active learning where the goal is to reduce the data annotation burden by interacting with a (human) oracle during training. Standard active learning methods ask the oracle to annotate data samples. Instead, we take a profoundly different approach: we ask for annotations of the decision boundary. We achieve this using a deep generative model to create novel instances along a 1d vector. A point on the decision boundary is revealed where the instances change class. Experimentally we show on three datasets that our method can be plugged-in to other active learning schemes, that human oracles can effectively annotate point on the decision boundary, and that decision boundary annotations improve ove r single sample instance annotations.

****

Convolutional Dictionary Learning via Local Processing

Vardan Papyan, Yaniv Romano, Jeremias Sulam, Michael Elad; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5296-5304

Convolutional Sparse Coding (CSC) is an increasingly popular model in the signal and image processing communities, tackling some of the limitations of traditional patch-based sparse representations. Although several works have addressed the dictionary learning problem under this model, these relied on an ADMM formulation in the Fourier domain, losing the sense of locality and the relation to the traditional patch-based sparse pursuit. A recent work suggested a novel theoretical analysis of this global model, providing guarantees that rely on a localized sparsity measure. Herein, we extend this local-global relation by showing how one can efficiently solve the convolutional sparse pursuit problem and train the filters involved, while operating locally on image patches. Our approach provides an intuitive algorithm that can leverage standard techniques from the sparse representations field. The proposed method is fast to train, simple to implement, and flexible enough that it can be easily deployed in a variety of applications. We demonstrate the proposed training scheme for image inpainting and image separation, while achieving state-of-the-art results.

****

Editable Parametric Dense Foliage From 3D Capture

Gaurav Chaurasia, Paul Beardsley; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5305-5314

We present an algorithm to compute parametric models of dense foliage. The guiding principles of our work are automatic reconstruction and compact artist friendly representation. We use Bezier patches to model leaf surface, which we compute from images and point clouds of dense foliage. We present an algorithm to segment individual leaves from colour and depth data. We then reconstruct the Bezier representation from segmented leaf points clouds using non-linear optimisation. Unlike previous work, we do not require laboratory scanned exemplars or user intervention. We also demonstrate intuitive manipulators to edit the reconstructed parametric models. We believe our work is a step towards making captured data more accessible to artists for foliage modelling.

****

Refractive Structure-From-Motion Through a Flat Refractive Interface

Francois Chadebecq, Francisco Vasconcelos, George Dwyer, Rene Lacher, Sebastien Ourselin, Tom Vercauteren, Danail Stoyanov; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5315-5323

Recovering 3D scene geometry from underwater images involves the Refractive Structure-from-Motion (RSfM) problem, where the image distortions caused by light refraction at the interface between different propagation media invalidates the single view point assumption. Direct use of the pinhole camera model in RSfM leads to inaccurate camera pose estimation and consequently drift. RSfM methods have been thoroughly studied for the case of a thick glass interface that assumes two refractive interfaces between the camera and the viewed scene. On the other hand, when the camera lens is in direct contact with the water, there is only one refractive interface. By explicitly considering a refractive interface, we develop a succinct derivation of the refractive fundamental matrix in the form of the generalised epipolar constraint for an axial camera. We use the refractive fundamental matrix to refine initial pose estimates obtained by assuming the pinhole model. This strategy allows us to robustly estimate underwater camera poses, where other methods suffer from poor noise-sensitivity. We also formulate a new four view constraint enforcing camera pose consistency along a video which leads us to a novel RSfM framework. For validation we use synthetic data to show the numerical properties of our method and we provide results on real data to demonstrate performance within laboratory settings and for applications in endoscopy.

****

Submodular Trajectory Optimization for Aerial 3D Scanning

Mike Roberts, Debadeepta Dey, Anh Truong, Sudipta Sinha, Shital Shah, Ashish Kap oor, Pat Hanrahan, Neel Joshi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5324-5333

Drones equipped with cameras are emerging as a powerful tool for large-scale aerial 3D scanning, but existing automatic flight planners do not exploit all available information about the scene, and can therefore produce inaccurate and incomplete 3D models. We present an automatic method to generate drone trajectories, such that the imagery acquired during the flight will later produce a high-fidelity 3D model. Our method uses a coarse estimate of the scene geometry to plan camera trajectories that: (1) cover the scene as thoroughly as possible; (2) encourage observations of scene geometry from a diverse set of viewing angles; (3) avoid obstacles; and (4) respect a user-specified flight time budget. Our method relies on a mathematical model of scene coverage that exhibits an intuitive diminishing returns property known as submodularity. We leverage this property extensively to design a trajectory planning algorithm that reasons globally about the non-additive coverage reward obtained across a trajectory, jointly with the cost of traveling between views. We evaluate our method by using it to scan three large outdoor scenes, and we perform a quantitative evaluation using a photorealistic video game simulator.

****

Camera Calibration by Global Constraints on the Motion of Silhouettes

Gil Ben-Artzi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5334-5343

We address the problem of epipolar geometry using the motion of silhouettes. Such methods match epipolar lines or frontier points across views, which are then used as the set of putative correspondences. We introduce an approach that improves by two orders of magnitude the performance over state-of-the-art methods, by significantly reducing the number of outliers in the putative matching. We model the frontier points' correspondence problem as constrained flow optimization, requiring small differences between their coordinates over consecutive frames. Our approach is formulated as a Linear Integer Program and we show that due to the nature of our problem, it can be solved efficiently in an iterative manner. Our method was validated on four standard datasets providing accurate calibrations across very different viewpoints.

****

Deltille Grids for Geometric Camera Calibration

Hyowon Ha, Michal Perdoch, Hatem Alismail, In So Kweon, Yaser Sheikh; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5344-5352

The recent proliferation of high resolution cameras presents an opportunity to achieve unprecedented levels of precision in visual 3D reconstruction. Yet the camera calibration pipeline, developed decades ago using checkerboards, has remained the de facto standard. In this paper, we ask the question: are checkerboards the optimal pattern for high precision calibration? We empirically demonstrate that deltille grids (regular triangular tiling) produce the highest precision calibration of the possible tilings of Euclidean plane. We posit that they should be the new standard for high-precision calibration and present a complete ecosystem for calibration using deltille grids including: (1) a highly precise corner detection algorithm based on polynomial surface fitting; (2) an indexing scheme based on polarities extracted from the fitted surfaces; and (3) a 2D coding system for deltille grids, which we refer to as DelTags, in lieu of conventional matrix barcodes. We demonstrate state-of-the-art performance and apply the full calibration ecosystem through the use of 3D calibration objects for multiview camera calibration.

****

A Lightweight Single-Camera Polarization Compass With Covariance Estimation

Wolfgang Sturzl; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5353-5361

A lightweight visual compass system is presented as well as a direct method for estimating sun direction and its covariance. The optical elements of the system are described enabling estimation of sky polarization in a FOV of approx. 56 deg

rees with a single standard camera sensor. Using the proposed direct method, the sun direction and its covariance matrix can be estimated based on the polarization measured in the image plane. Experiments prove the applicability of the polarization sensor and the proposed estimation method, even in difficult conditions. It is also shown that in case the sensor is not leveled, combination with an IMU allows to determine all degrees of orientation. Due to the low weight of the sensor and the low complexity of the estimation method the polarization system is well suited for MAVs which have limited payload and computational resources. Furthermore, since not just the sun direction but also its covariance is estimated an integration in a multi-sensor navigation framework is straight forward.
****
Reflectance Capture Using Univariate Sampling of BRDFs
Zhuo Hui, Kalyan Sunkavalli, Joon-Young Lee, Sunil Hadap, Jian Wang, Aswin C. Sankaranarayanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5362-5370
We propose the use of a light-weight setup consisting of a collocated camera and light source --- commonly found on mobile devices --- to reconstruct surface normals and spatially-varying BRDFs of near-planar material samples. A collocated setup provides only a 1-D "univariate" sampling of the 4-D BRDF. We show that a univariate sampling is sufficient to estimate parameters of commonly used analytical BRDF models. Subsequently, we use a dictionary-based reflectance prior to derive a robust technique for per-pixel normal and BRDF estimation. We demonstrate real-world shape and capture, and its application to material editing and classification, using real data acquired using a mobile phone.
****
Estimating Defocus Blur via Rank of Local Patches
Guodong Xu, Yuhui Quan, Hui Ji; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5371-5379
This paper addresses the problem of defocus map estimation from a single image. We present a fast yet effective approach to estimate the spatially varying amounts of defocus blur at edge locations, which is based on the maximum, ranks of the corresponding local patches with different orientations in gradient domain. Such an approach is motivated by the theoretical analysis which reveals the connection between the rank of a local patch blurred by a defocus blur kernel and the blur amount by the kernel. After the amounts of defocus blur at edge locations are obtained, a complete defocus map is generated by a standard propagation procedure. The proposed method is extensively evaluated on real image datasets, and the experimental results show its superior performance to existing approaches.proposed method is extensively evaluated on real data, and the experimental results show its superior performance to existing approaches.
****
RGB-Infrared Cross-Modality Person Re-Identification
Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, Jianhuang Lai; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5380-5389
Person re-identification (Re-ID) is an important problem in video surveillance, aiming to match pedestrian images across camera views. Currently, most works focus on RGB-based Re-ID. However, in some applications, RGB images are not suitable, e.g. in a dark environment or at night. Infrared (IR) imaging becomes necessary in many visual systems. To that end, matching RGB images with infrared images is required, which are heterogeneous with very different visual characteristics. For person Re-ID, this is a very challenging cross-modality problem that has not been studied so far. In this work, we address the RGB-IR cross-modality Re-ID problem and contribute a new multiple modality Re-ID dataset named SYSU-MM01, including RGB and IR images of 491 identities from 6 cameras, giving in total 287,628 RGB images and 15,792 IR images. To explore the RGB-IR Re-ID problem, we evaluate existing popular cross-domain models, including three commonly used neural network structures (one-stream, two-stream and asymmetric FC layer) and analyse the relation between them. We further propose deep zero-padding for training one-stream network towards automatically evolving domain-specific nodes in the ne

twork for cross-modality matching. Our experiments show that RGB-IR cross-modality matching is very challenging but still feasible using the proposed model with deep zero-padding, giving the best performance. Our dataset is available at http://isee.sysu.edu.cn/project/RGBIRReID.htm.

****

Intrinsic 3D Dynamic Surface Tracking Based on Dynamic Ricci Flow and Teichmuller Map

Xiaokang Yu, Na Lei, Yalin Wang, Xianfeng Gu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5390-5398

3D dynamic surface tracking is an important research problem and plays a vital role in many computer vision and medical imaging applications. However, it is still challenging to efficiently register surface sequences which has large deformations and strong noise. In this paper, we propose a novel automatic method for non-rigid 3D dynamic surface tracking with surface Ricci flow and Teichmuller map methods. According to quasi-conformal Teichmuller theory, the Techmuller map minimizes the maximal dilation so that our method is able to automatically register surfaces with large deformations. Besides, the adoption of Delaunay triangulation and quadrilateral meshes makes our method applicable to low quality meshes. In our work, the 3D dynamic surfaces are acquired by a high speed 3D scanner. We first identified sparse surface features using machine learning methods in the texture space. Then we assign landmark features with different curvature settings and the Riemannian metric of the surface is computed by the dynamic Ricci flow method, such that all the curvatures are concentrated on the feature points and the surface is flat everywhere else. The registration among frames is computed by the Teichmuller mappings, which aligns the feature points with least angle distortions. We apply our new method to multiple sequences of 3D facial surfaces with large expression deformations and compare them with two other state-of-the-art tracking methods. The effectiveness of our method is demonstrated by the clearly improved accuracy and efficiency.

****

Multi-Scale Deep Learning Architectures for Person Re-Identification

Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, Xiangyang Xue; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5399-5408

Person Re-identification (re-id) aims to match people across non-overlapping camera views in a public space. It is a challenging problem because many people captured in surveillance videos wear similar clothes. Consequently, the differences in their appearance are often subtle and only detectable at the right location and scales. Existing re-id models, particularly the recently proposed deep learning based ones match people at a single scale. In contrast, in this paper, a novel multi-scale deep learning model is proposed. Our model is able to learn deep discriminative feature representations at different scales and automatically determine the most suitable scales for matching. The importance of different spatial locations for extracting discriminative features is also learned explicitly. Experiments are carried out to demonstrate that the proposed model outperforms the state-of-the art on a number of benchmarks.

****

Range Loss for Deep Face Recognition With Long-Tailed Training Data

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, Yu Qiao; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5409-5418

Deep convolutional neural networks have achieved significant improvements on face recognition task due to their ability to learn highly discriminative features from tremendous amounts of face images. Many large scale face datasets exhibit long-tail distribution where a small number of entities (persons) have large number of face images while a large number of persons only have very few face samples (long tail). Most of the existing works alleviate this problem by simply cutting the tailed data and only keep identities with enough number of examples. Unlike these work, this paper investigated how long-tailed data impact the training of face CNNs and develop a novel loss function, called range loss, to effectively utilize the tailed data in training process. More specifically, range loss is designed to reduce overall intrapersonal variations while enlarge inter-personal differences simultaneously. Extensive experiments on two face recognition benchmarks, Labeled Faces in the Wild (LFW) and YouTube Faces (YTF), demonstrate the effectiveness of the proposed range loss in overcoming the long tail effect, and show the good generalization ability of the proposed methods.

****

Face Sketch Matching via Coupled Deep Transform Learning

Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, Afzel Noore, Angshul Majumdar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5419-5428

Face sketch to digital image matching is an important challenge of face recognition that involves matching across different domains. Current research efforts have primarily focused on extracting domain invariant representations or learning a mapping from one domain to the other. In this research, we propose a novel transform learning based approach termed as DeepTransformer, which learns a transformation and mapping function between the features of two domains. The proposed formulation is independent of the input information and can be applied with any existing learned or hand-crafted feature. Since the mapping function is directional in nature, we propose two variants of DeepTransformer: (i) semi-coupled and (ii) symmetrically-coupled deep transform learning. This research also uses a novel IIIT-D Composite Sketch with Age (CSA) variations database which contains sketch images of 150 subjects along with age-separated digital photos. The performance of the proposed models is evaluated on a novel application of sketch-to-sketch matching, along with sketch-to-digital photo matching. Experimental results demonstrate the robustness of the proposed models in comparison to existing state-of-the-art sketch matching algorithms and a commercial face recognition system.

****

Realistic Dynamic Facial Textures From a Single Image Using GANs

Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, Hao Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5429-5438

We present a novel method to realistically puppeteer and animate a face from a single RGB image using a source video sequence. We begin by fitting a multilinear PCA model to obtain the 3D geometry and a single texture of the target face. In order for the animation to be realistic, however, we need dynamic per-frame textures that capture subtle wrinkles and deformations corresponding to the animated facial expressions. This problem is highly underconstrained, as dynamic textures cannot be obtained directly from a single image. Furthermore, if the target face has a closed mouth, it is not possible to obtain actual images of the mouth interior. To address this issue, we train a Deep Generative Network that can infer realistic per-frame texture deformations, including the mouth interior, of the target identity using the per-frame source textures and the single target texture. By retargeting the PCA expression geometry from the source, as well as using the newly inferred texture, we can both animate the face and perform video face replacement on the source video using the target appearance.

****

Pixel Recursive Super Resolution

Ryan Dahl, Mohammad Norouzi, Jonathon Shlens; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5439-5448

Super resolution is the problem of artificially enlarging a low resolution photograph to recover a plausible high resolution version. In the regime of high magnification factors, the problem is dramatically underspecified and many plausible, high resolution images may match a given low resolution image. In particular, traditional super resolution techniques fail in this regime due to the multimodality of the problem and strong prior information that must be imposed on image synthesis to produce plausible high resolution images. In this work we propose a new probabilistic deep network architecture, a pixel recursive super resolution model, that is an extension of PixelCNNs to address this problem. We demonstrate that this model produces a diversity of plausible high resolution images at large magnification factors. Furthermore, in human evaluation studies we demonstrate how previous methods fail to fool human observers. However, high resolution im ages sampled from this probabilistic deep network do fool a naive human observer a significant fraction of the time.

****

PanNet: A Deep Network Architecture for Pan-Sharpening

Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, John Paisley; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5449-5457

We propose a deep network architecture for the pan-sharpening problem called PanNet. We incorporate domain-specific knowledge to design our PanNet architecture by focusing on the two aims of the pan-sharpening problem: spectral and spatial preservation. For spectral preservation, we add up-sampled multispectral images to the network output, which directly propagates the spectral information to the reconstructed image. To preserve spatial structure, we train our network parameters in the high-pass filtering domain rather than the image domain. We show that the trained network generalizes well to images from different satellites without needing retraining. Experiments show significant improvement over state-of-the-art methods visually and in terms of standard quality metrics.

****

Recurrent Color Constancy

Yanlin Qian, Ke Chen, Jarno Nikkanen, Joni-Kristian Kamarainen, Jiri Matas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5458-5466

We introduce a novel formulation of temporal color constancy which considers multiple frames preceding the frame for which illumination is estimated. We propose an end-to-end trainable recurrent color constancy network -- the RCC-Net -- which exploits convolutional LSTMs and a simulated sequence to learn compositional representations in space and time. We use a standard single frame color constancy benchmark, the SFU Gray Ball Dataset, which can be adapted to a temporal setting. Extensive experiments show that the proposed method consistently outperforms single-frame state-of-the-art methods and their temporal variants.

****

Saliency Pattern Detection by Ranking Structured Trees

Lei Zhu, Haibin Ling, Jin Wu, Huiping Deng, Jin Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5467-5476

In this paper we propose a new salient object detection method via structured label prediction. By learning appearance features in rectangular regions, our structural region representation encodes the local saliency distribution with a matrix of binary labels. We show that the linear combination of structured labels can well model the saliency distribution in local regions. Representing region saliency with structured labels has two advantages: 1) it connects the label assignment of all enclosed pixels, which produces a smooth saliency prediction; and 2) regular-shaped nature of structured labels enables well definition of traditional cues such as regional properties and center surround contrast, and these cues help to build meaningful and informative saliency measures. To measure the consistency between a structured label and the corresponding saliency distribution, we further propose an adaptive label ranking algorithm using proposals that are generated by a CNN model. Finally, we introduce a K-NN enhanced graph representation for saliency propagation, which is more favorable for our task than the widely-used adjacent-graph-based ones. Experimental results demonstrate the effectiveness of our proposed method on six popular benchmarks compared with state-of-the-art approaches.

****

Monocular Video-Based Trailer Coupler Detection Using Multiplexer Convolutional Neural Network

Yousef Atoum, Joseph Roth, Michael Bliss, Wende Zhang, Xiaoming Liu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5477-5485

This paper presents an automated monocular-camera-based computer vision system for autonomous self-backing-up a vehicle towards a trailer, by continuously estimating the 3D trailer coupler position and feeding it to the vehicle control syst em, until the alignment of the tow hitch with the trailers coupler. This system is made possible through our proposed distance-driven Multiplexer-CNN method, which selects the most suitable CNN using the estimated coupler-to-vehicle distance. The input of the multiplexer is a group made of a CNN detector, trackers, and 3D localizer. In the CNN detector, we propose a novel algorithm to provide a presence confidence score with each detection. The score reflects the existence of the target object in a region, as well as how accurate is the 2D target detection. We demonstrate the accuracy and efficiency of the system on a large trailer database. Our system achieves an estimation error of 1.4 cm when the ball reaches the coupler, while running at 18.9 FPS on a regular PC.

****

Parallel Tracking and Verifying: A Framework for Real-Time and High Accuracy Visual Tracking

Heng Fan, Haibin Ling; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5486-5494

Being intensively studied, visual tracking has seen great recent advances in either speed (e.g., with correlation filters) or accuracy (e.g., with deep features). Real-time and high accuracy tracking algorithms, however, remain scarce. In this paper we study the problem from a new perspective and present a novel parallel tracking and verifying (PTAV) framework, by taking advantage of the ubiquity of multi-thread techniques and borrowing from the success of parallel tracking and mapping in visual SLAM. Our PTAV framework typically consists of two components, a tracker T and a verifier V, working in parallel on two separate threads. The tracker T aims to provide a super real-time tracking inference and is expected to perform well most of the time; by contrast, the verifier V checks the tracking results and corrects T when needed. The key innovation is that, V does not work on every frame but only upon the requests from T; on the other end, T may adjust the tracking according to the feedback from V. With such collaboration, PTAV enjoys both the high efficiency provided by T and the strong discriminative power by V. In our extensive experiments on popular benchmarks including OTB2013, OTB2015, TC128 and UAV20L, PTAV achieves the best tracking accuracy among all real-time trackers, and in fact performs even better than many deep learning based solutions. Moreover, as a general framework, PTAV is very flexible and has great rooms for improvement and generalization.

****

Non-Rigid Object Tracking via Deformable Patches Using Shape-Preserved KCF and Level Sets

Xin Sun, Ngai-Man Cheung, Hongxun Yao, Yiluan Guo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5495-5503

Part-based trackers are effective in exploiting local details of the target object for robust tracking. In contrast to most existing part-based methods that divide all kinds of target objects into a number of fixed rectangular patches, in this paper, we propose a novel framework in which a set of deformable patches dynamically collaborate on tracking of non-rigid objects. In particular, we proposed a shape-preserved kernelized correlation filter (SP-KCF) which can accommodate target shape information for robust tracking. The SP-KCF is introduced into the level set framework for dynamic tracking of individual patches. In this manner, our proposed deformable patches are target-dependent, have the capability to assume complex topology, and are deformable to adapt to target variations. As these deformable patches properly capture individual target subregions, we exploit their photometric discrimination and shape variation to reveal the trackability of individual target subregions, which enables the proposed tracker to dynamically take advantage of those subregions with good trackability for target likelihood estimation. Finally the shape information of these deformable patches enables accurate object contours to be computed as the tracking output. Experimental results on the latest public sets of challenging sequences demonstrate the effectiveness of the proposed method.

****

A Discriminative View of MRF Pre-Processing Algorithms

Chen Wang, Charles Herrmann, Ramin Zabih; Proceedings of the IEEE International

Conference on Computer Vision (ICCV), 2017, pp. 5504-5513

While Markov Random Fields (MRFs) are widely used in computer vision, they present a quite challenging inference problem. MRF inference can be accelerated by pre-processing techniques like Dead End Elimination (DEE) or QPBO-based approaches which compute the optimal labeling of a subset of variables. These techniques are guaranteed to never wrongly label a variable but they often leave a large number of variables unlabeled. We address this shortcoming by interpreting pre-processing as a classification problem, which allows us to trade off false positives (i.e., giving a variable an incorrect label) versus false negatives (i.e., failing to label a variable). We describe an efficient discriminative rule that finds optimal solutions for a subset of variables. Our technique provides both per-instance and worst-case guarantees concerning the quality of the solution. Empirical studies were conducted over several benchmark datasets. We obtain a speedup factor of 2 to 12 over expansion moves without preprocessing, and on difficult non-submodular energy functions produce slightly lower energy.

****

Offline Handwritten Signature Modeling and Verification Based on Archetypal Analysis

Elias N. Zois, Ilias Theodorakopoulos, George Economou; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5514-5523

The handwritten signature is perhaps the most accustomed way for the acknowledgement of the consent of an individual or the authentication of the identity of a person in numerous transactions. In addition, the authenticity of a questioned offline or static handwritten signature still poses a case of interest, especially in forensic related applications. A common approach in offline signature verification system is to apply several predetermined image analysis models. Consequently, any offline signature sample which originates from either authentic persons or forgers, utilizes a fixed feature extraction base. In this proposed study, the feature space and the corresponding projection values depend on the training samples only; thus the proposed method can be found useful in forensic cases. In order to do so, we reenter a groundbreaking unsupervised learning method named archetypal analysis, which is connected to effective data analysis approaches such as sparse coding. Due to the fact that until recently there was no efficient implementation publicly available, archetypal analysis had only few cases of use. However, a fast optimization scheme using an active set strategy is now available. The main goal of this work is to introduce archetypal analysis for offline signature verification. The output of the archetypal analysis of few reference samples is a set of archetypes which are used to form the base of the feature space. Then, given a set of archetypes and a signature sample under examination archetypal analysis and average pooling provides the corresponding features. The promising performance of the proposed approach is demonstrated with the use of an evaluation method which employs the popular CEDAR and MCYT75 signature datasets .

****

Long Short-Term Memory Kalman Filters: Recurrent Neural Estimators for Pose Regularization

Huseyin Coskun, Felix Achilles, Robert DiPietro, Nassir Navab, Federico Tombari; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5524-5532

One-shot pose estimation for tasks such as body joint localization, camera pose estimation, and object tracking are generally noisy, and temporal filters have been extensively used for regularization. One of the most widely-used methods is the Kalman filter, which is both extremely simple and general. However, Kalman filters require a motion model and measurement model to be specified a priori, which burdens the modeler and simultaneously demands that we use explicit models that are often only crude approximations of reality. For example, in the pose-estimation tasks mentioned above, it is common to use motion models that assume constant velocity or constant acceleration, and we believe that these simplified representations are severely inhibitive. In this work, we propose to instead learn rich, dynamic representations of the motion and noise models. In particular, we propose learning these models from data using long short-term memory, which allows representations that depend on all previous observations and all previous states. We evaluate our method using three of the most popular pose estimation tasks in computer vision, and in all cases we obtain state-of-the-art performance.

****

Learning Spatio-Temporal Representation With Pseudo-3D Residual Networks

Zhaofan Qiu, Ting Yao, Tao Mei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5533-5541

Convolutional Neural Networks (CNN) have been regarded as a powerful class of models for image recognition problems. Nevertheless, it is not trivial when utilizing a CNN for learning spatio-temporal video representation. A few studies have shown that performing 3D convolutions is a rewarding approach to capture both spatial and temporal dimensions in videos. However, the development of a very deep 3D CNN from scratch results in expensive computational cost and memory demand. A valid question is why not recycle off-the-shelf 2D networks for a 3D CNN. In this paper, we devise multiple variants of bottleneck building blocks in a residual learning framework by simulating 3*3*3 convolutions with 1*3*3 convolutional filters on spatial domain (equivalent to 2D CNN) plus 3*1*1 convolutions to construct temporal connections on adjacent feature maps in time. Furthermore, we propose a new architecture, named Pseudo-3D Residual Net (P3D ResNet), that exploits all the variants of blocks but composes each in different placement of ResNet, following the philosophy that enhancing structural diversity with going deep could improve the power of neural networks. Our P3D ResNet achieves clear improvements on Sports-1M video classification dataset against 3D CNN and frame-based 2D CNN by 5.3% and 1.8%, respectively. We further examine the generalization performance of video representation produced by our pre-trained P3D ResNet on five different benchmarks and three different tasks, demonstrating superior performances over several state-of-the-art techniques.

****

Deeper, Broader and Artier Domain Generalization

Da Li, Yongxin Yang, Yi-Zhe Song, Timothy M. Hospedales; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5542-5550

The problem of domain generalization is to learn from multiple training domains, and extract a domain-agnostic model that can then be applied to an unseen domain. Domain generalization (DG) has a clear motivation in contexts where there are target domains with distinct characteristics, yet sparse data for training. For example recognition in sketch images, which are distinctly more abstract and rarer than photos. Nevertheless, DG methods have primarily been evaluated on photo-only benchmarks focusing on alleviating the dataset bias where both problems of domain distinctiveness and data sparsity can be minimal. We argue that these benchmarks are overly straightforward, and show that simple deep learning baselines perform surprisingly well on them. In this paper, we make two main contributions: Firstly, we build upon the favorable domain shift-robust properties of deep learning methods, and develop a low-rank parameterized CNN model for end-to-end DG learning. Secondly, we develop a DG benchmark dataset covering photo, sketch, cartoon and painting domains. This is both more practically relevant, and harder (bigger domain shift) than existing benchmarks. The results show that our method outperforms existing DG alternatives, and our dataset provides a more significant DG challenge to drive future research.

****

Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval

Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5551-5560

Human sketches are unique in being able to capture both the spatial topology of a visual object, as well as its subtle appearance details. Fine-grained sketch-based image retrieval (FG-SBIR) importantly leverages on such fine-grained characteristics of sketches to conduct instance-level retrieval of photos. Nevertheless, human sketches are often highly abstract and iconic, resulting in severe misalignments with candidate photos which in turn make subtle visual detail matching

difficult. Existing FG-SBIR approaches focus only on coarse holistic matching via deep cross-domain representation learning, yet ignore explicitly accounting for fine-grained details and their spatial context. In this paper, a novel deep FG-SBIR model is proposed which differs significantly from the existing models in that: (1) It is spatially aware, achieved by introducing an attention module that is sensitive to the spatial position of visual details; (2) It combines coarse and fine semantic information via a shortcut connection fusion block; and (3) It models feature correlation and is robust to misalignments between the extracted features across the two domains by introducing a novel higher order learnable energy function (HOLEF) based loss. Extensive experiments show that the proposed deep spatial-semantic attention model significantly outperforms the state-of-the-art.
****
Soft-NMS -- Improving Object Detection With One Line of Code
Navaneeth Bodla, Bharat Singh, Rama Chellappa, Larry S. Davis; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5561-5569
Non-maximum suppression is an integral part of the object detection pipeline. First, it sorts all detection boxes on the basis of their scores. The detection box M with the maximum score is selected and all other detection boxes with a significant overlap (using a pre-defined threshold) with M are suppressed. This process is recursively applied on the remaining boxes. As per the design of the algorithm, if an object lies within the predefined overlap threshold, it leads to a miss. To this end, we propose Soft-NMS, an algorithm which decays the detection scores of all other objects as a continuous function of their overlap with M. Hence, no object is eliminated in this process. Soft-NMS obtains consistent improvements for the coco-style mAP metric on standard datasets like PASCAL VOC 2007 (1.7% for both R-FCN and Faster-RCNN) and MS-COCO (1.3% for R-FCN and 1.1% for Faster-RCNN) by just changing the NMS algorithm without any additional hyper-parameters. Using Deformable-RFCN, Soft-NMS improves state-of-the-art in object detection from 39.8% to 40.9% with a single model. Further, the computational complexity of Soft-NMS is the same as traditional NMS and hence it can be efficiently implemented. Since Soft-NMS does not require any extra training and is simple to implement, it can be easily integrated into any object detection pipeline. Code for Soft-NMS is publicly available on GitHub
****
Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images
Aron Yu, Kristen Grauman; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5570-5579
Distinguishing subtle differences in attributes is valuable, yet learning to make visual comparisons remains nontrivial. Not only is the number of possible comparisons quadratic in the number of training images, but also access to images adequately spanning the space of fine-grained visual differences is limited. We propose to overcome the sparsity of supervision problem via synthetically generated images. Building on a state-of-the-art image generation engine, we sample pairs of training images exhibiting slight modifications of individual attributes. Augmenting real training image pairs with these examples, we then train attribute ranking models to predict the relative strength of an attribute in novel pairs of real images. Our results on datasets of faces and fashion images show the great promise of bootstrapping imperfect image generators to counteract sample sparsity for learning to rank.
****
Video Scene Parsing With Predictive Feature Learning
Xiaojie Jin, Xin Li, Huaxin Xiao, Xiaohui Shen, Zhe Lin, Jimei Yang, Yunpeng Chen, Jian Dong, Luoqi Liu, Zequn Jie, Jiashi Feng, Shuicheng Yan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5580-5588
Video scene parsing is challenging due to the following two reasons: firstly, it is non-trivial to learn meaningful video representations for producing the temporally consistent labeling map; secondly, such a learning process becomes more difficult with insufficient labeled video training data. In this work, we propose a unified framework to address the above two problems, which is to our knowledg

e the first model to employ predictive feature learning in the video scene parsing. The predictive feature learning is carried out in two predictive tasks: frame prediction and predictive parsing. It is experimentally proved that the learned predictive features in our model are able to significantly enhance the video parsing performance by combining with the standard image parsing network. Interestingly, the performance gain brought by the predictive learning is almost costless as the features are learned from a large amount of unlabeled video data in an unsupervised way. Extensive experiments over two challenging datasets, Cityscapes and Camvid, have demonstrated the effectiveness of our model by showing remarkable improvement over well-established baselines.

****

Understanding and Mapping Natural Beauty

Scott Workman, Richard Souvenir, Nathan Jacobs; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5589-5598

While natural beauty is often considered a subjective property of images, in this paper, we take an objective approach and provide methods for quantifying and predicting the scenicness of an image. Using a dataset containing hundreds of thousands of outdoor images captured throughout Great Britain with crowdsourced ratings of natural beauty, we propose an approach to predict scenicness which explicitly accounts for the variance of human ratings. We demonstrate that quantitative measures of scenicness can benefit semantic image understanding, content-aware image processing, and a novel application of cross-view mapping, where the sparsity of ground-level images can be addressed by incorporating unlabeled overhead images in the training and prediction steps. For each application, our methods for scenicness prediction result in quantitative and qualitative improvements over baseline approaches.

****

Human Pose Estimation Using Global and Local Normalization

Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, Jingdong Wang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5599-5607

In this paper, we address the problem of estimating the positions of human joints, i.e., articulated pose estimation. Recent state-of-the-art solutions model two key issues, joint detection and spatial configuration refinement, together using convolutional neural networks. Our work mainly focuses on spatial configuration refinement by reducing variations of human poses statistically, which is motivated by the observation that the scattered distribution of the relative locations of joints (e.g., the left wrist is distributed nearly uniformly in a circular area around the left shoulder) makes the learning of convolutional spatial models hard. We present a two-stage normalization scheme, human body normalization and limb normalization, to make the distribution of the relative joint locations compact, resulting in easier learning of convolutional spatial models and more accurate pose estimation. In addition, our empirical results show that incorporating multi-scale supervision and multi-scale fusion into the joint detection network is beneficial. Experiment results demonstrate that our method consistently outperforms state-of-the-art methods on the benchmarks.

****

HashNet: Deep Learning to Hash by Continuation

Zhangjie Cao, Mingsheng Long, Jianmin Wang, Philip S. Yu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5608-5617

Learning to hash has been widely applied to approximate nearest neighbor search for large-scale multimedia retrieval, due to its computation efficiency and retrieval quality. Deep learning to hash, which improves retrieval quality by end-to-end representation learning and hash encoding, has received increasing attention recently. Subject to the ill-posed gradient difficulty in the optimization with sign activations, existing deep learning to hash methods need to first learn continuous representations and then generate binary hash codes in a separated binarization step, which suffer from substantial loss of retrieval quality. This work presents HashNet, a novel deep architecture for deep learning to hash by continuation method with convergence guarantees, which learns exactly binary hash co des from imbalanced similarity data. The key idea is to attack the ill-posed gradient problem in optimizing deep networks with non-smooth binary activations by continuation method, in which we begin from learning an easier network with smoothed activation function and let it evolve during the training, until it eventually goes back to being the original, difficult to optimize, deep network with the sign activation function. Comprehensive empirical evidence shows that HashNet can generate exactly binary hash codes and yield state-of-the-art multimedia retrieval performance on standard benchmarks.

****

Scaling the Scattering Transform: Deep Hybrid Networks

Edouard Oyallon, Eugene Belilovsky, Sergey Zagoruyko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5618-5627

We use the scattering network as a generic and fixed initialization of the first layers of a supervised hybrid deep network. We show that early layers do not necessarily need to be learned, providing the best results to-date with pre-defined representations while being competitive with Deep CNNs. Using a shallow cascade of 1x1 convolutions, which encodes scattering coefficients that correspond to spatial windows of very small sizes, permits to obtain AlexNet accuracy on the imagenet ILSVRC2012. We demonstrate that this local encoding explicitly learns invariance w.r.t. rotations. Combining scattering networks with a modern ResNet, we achieve a single-crop top 5 error of 11.4% on imagenet ILSVRC2012, comparable to the Resnet-18 architecture, while utilizing only 10 layers. We also find that hybrid architectures can yield excellent performance in the small sample regime, exceeding their end-to-end counterparts, through their ability to incorporate geometrical priors. We demonstrate this on subsets of the CIFAR-10 dataset and on the STL-10 dataset.

****

Flip-Invariant Motion Representation

Takumi Kobayashi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5628-5637

In action recognition, local motion descriptors contribute to effectively representing video sequences where target actions appear in localized spatio-temporal regions. For robust recognition, those fundamental descriptors are required to be invariant against horizontal (mirror) flipping in video frames which frequently occurs due to changes of camera viewpoints and action directions, deteriorating classification performance. In this paper, we propose methods to render flip invariance to the local motion descriptors by two approaches. One method leverages local motion flows to ensure the invariance on input patches where the descriptors are computed. The other derives a invariant form theoretically from the flipping transformation applied to hand-crafted descriptors. The method is also extended so as to deal with ConvNet descriptors through learning the invariant form based on data. The experimental results on human action classification show that the proposed methods favorably improve performance both of the handcrafted and the ConvNet descriptors.

****

Scene Categorization With Spectral Features

Salman H. Khan, Munawar Hayat, Fatih Porikli; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5638-5648

Spectral signatures of natural scenes were earlier found to be distinctive for different scene types with varying spatial envelope properties such as openness, naturalness, ruggedness, and symmetry. Recently, such handcrafted features have been outclassed by deep learning based representations. This paper proposes a novel spectral description of convolution features, implemented efficiently as a unitary transformation within deep network architectures. To the best of our knowledge, this is the first attempt to use deep learning based spectral features explicitly for image classification task. We show that the spectral transformation decorrelates convolutional activations, which reduces co-adaptation between feature detections, thus acts as an effective regularizer. Our approach achieves significant improvements on three large-scale scene-centric datasets (MIT-67, SUN-397, and Places-205). Furthermore, we evaluated the proposed approach on the att ribute detection task where its superior performance manifests its relevance to semantically meaningful characteristics of natural scenes.

****

Image2song: Song Retrieval via Bridging Image Content and Lyric Words

Xuelong Li, Di Hu, Xiaoqiang Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5649-5658

Image is usually taken for expressing some kinds of emotions or purposes, such as love, celebrating Christmas. There is another better way that combines the image and relevant song to amplify the expression, which has drawn much attention in the social network recently. Hence, the automatic selection of songs should be expected. In this paper, we propose to retrieve semantic relevant songs just by an image query, which is named as the image2song problem. Motivated by the requirements of establishing correlation in semantic/content, we build a semantic-based song retrieval framework, which learns the correlation between image content and lyric words. This model uses a convolutional neural network to generate rich tags from image regions, a recurrent neural network to model lyric, and then establishes correlation via a multi-layer perceptron. To reduce the content gap between image and lyric, we propose to make the lyric modeling focus on the main image content via a tag attention. We collect a dataset from the social-sharing multimodal data to study the proposed problem, which consists of (image, music clip, lyric) triplets. We demonstrate that our proposed model shows noticeable results in the image2song retrieval task and provides suitable songs. Besides, the song2image task is also performed.

****

Deep Functional Maps: Structured Prediction for Dense Shape Correspondence

Or Litany, Tal Remez, Emanuele Rodola, Alex Bronstein, Michael Bronstein; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5659-5667

We introduce a new framework for learning dense correspondence between deformable 3D shapes. Existing learning based approaches model shape correspondence as a labelling problem, where each point of a query shape receives a label identifying a point on some reference domain; the correspondence is then constructed a posteriori by composing the label predictions of two input shapes. We propose a paradigm shift and design a structured prediction model in the space of functional maps, linear operators that provide a compact representation of the correspondence. We model the learning process via a deep residual network which takes dense descriptor fields defined on two shapes as input, and outputs a soft map between the two given objects. The resulting correspondence is shown to be accurate on several challenging benchmarks comprising multiple categories, synthetic models, real scans with acquisition artifacts, topological noise, and partiality.

****

Training Deep Networks to Be Spatially Sensitive

Nicholas Kolkin, Eli Shechtman, Gregory Shakhnarovich; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5668-5677

In many computer vision tasks, for example saliency prediction or semantic segmentation, the desired output is a foreground map that predicts pixels where some criteria is satisfied. Despite the inherently spatial nature of this task commonly used learning objectives do not incorporate the spatial relationships between misclassified pixels and the underlying ground truth. The Weighted F-measure, a recently proposed evaluation metric, does reweight errors spatially, and has been shown to closely correlate with human evaluation of quality, and stably rank predictions with respect to noisy ground truths (such as a sloppy human annotator might generate). However it suffers from computational complexity which makes it intractable as an optimization objective for gradient descent, which must be evaluated thousands or millions of times while learning a model's parameters. We propose a differentiable and efficient approximation of this metric. By incorporating spatial information into the objective we can use a simpler model than competing methods without sacrificing accuracy, resulting in faster inference speeds and alleviating the need for pre/post-processing. We match (or improve) performance on several tasks compared to prior state of the art by traditional metric s, and in many cases significantly improve performance by the weighted F-measure.

****

3DCNN-DQN-RNN: A Deep Reinforcement Learning Framework for Semantic Parsing of Large-Scale 3D Point Clouds

Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou, Rongtian Ye, Yuebin Wang, Jiwen Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5678-5687

Semantic parsing of large-scale 3D point clouds is an important research topic in computer vision and remote sensing fields. Most existing approaches utilize hand-crafted features for each modality independently and combine them in a heuristic manner. They often fail to consider the consistency and complementary information among features adequately, which makes them difficult to capture high-level semantic structures. The features learned by most of the current deep learning methods can obtain high-quality image classification results. However, these methods are hard to be applied to recognize 3D point clouds due to unorganized distribution and various point density of data. In this paper, we propose a 3DCNN-DQN-RNN method which fuses the 3D convolutional neural network (CNN), Deep Q-Network (DQN) and Residual recurrent neural network (RNN) for an efficient semantic parsing of large-scale 3D point clouds. In our method, an eye window under control of the 3D CNN and DQN can localize and segment the points of the class objects efficiently. The 3D CNN and Residual RNN further extract robust and discriminative features of the points in the eye window, and thus greatly enhance the parsing accuracy of large-scale point clouds. Our method provides an automatic process that maps the raw data to the classification results. It also integrates object localization, segmentation and classification into one framework. Experimental results demonstrate that the proposed method outperforms the state-of-the-art point cloud classification methods.

****

Semi Supervised Semantic Segmentation Using Generative Adversarial Network

Nasim Souly, Concetto Spampinato, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5688-5696

Semantic segmentation has been a long standing challenging task in computer vision. It aims at assigning a label to each image pixel and needs a significant number of pixel-level annotated data, which is often unavailable. To address this lack of annotations, in this paper, we leverage, on one hand, a massive amount of available unlabeled or weakly labeled data, and on the other hand, non-realimages created through Generative Adversarial Networks. In particular, we propose a semi-supervised framework -based on Generative Adversarial Networks (GANs) - which consists of a generator network to provide extra training examples to a multi-class classifier, acting as discriminator in the GAN framework, that assigns sample a label y from the K possible classes or marks it as a fake sample (extra class). The underlying idea is that adding large fake visual data forces real samples to be close in the feature space, which, in turn, improves multiclass pixel classification. To ensure a higher quality of generated images by GANs with consequently improved pixel classification, we extend the above framework by adding weakly annotated data, i.e., we provide class level information to the generator. We test our approaches on several challenging benchmarking visual datasets, i.e. PASCAL, SiftFLow, Stanford and CamVid, achieving competitive performance compared to state-of-the-art semantic segmentation methods.

****

Efficient Low Rank Tensor Ring Completion

Wenqi Wang, Vaneet Aggarwal, Shuchin Aeron; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5697-5705

Using the matrix product state (MPS) representation of the recently proposed tensor ring (TR) decompositions, in this paper we propose a TR completion algorithm, which is an alternating minimization algorithm that alternates over the factors in the MPS representation. This development is motivated in part by the success of matrix completion algorithms that alternate over the (low-rank) factors. We propose a novel initialization method and analyze the computational complexity of the TR completion algorithm. The numerical comparison between the TR completion algorithm and the existing algorithms that employ a low rank tensor train (TT) approximation for data completion shows that our method outperforms the existing ones for a variety of real computer vision settings, and thus demonstrates the improved expressive power of tensor ring as compared to tensor train.

****

Semantic Image Synthesis via Adversarial Learning

Hao Dong, Simiao Yu, Chao Wu, Yike Guo; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5706-5714

In this paper, we propose a way of synthesizing realistic images directly with natural language description, which has many useful applications, e.g.intelligent image manipulation. We attempt to accomplish such synthesis: given a source image and a target text description, our model synthesizes images to meet two requirements: 1) being realistic while matching the target text description; 2) maintaining other image features that are irrelevant to the text description. The model should be able to disentangle the semantic information from the two modalities (image and text), and generate new images from the combined semantics. To achieve this, we proposed an end-to-end neural architecture that leverages adversarial learning to automatically learn implicit loss functions, which are optimized to fulfill the aforementioned two requirements. We have evaluated our model by conducting experiments on Caltech-200 bird dataset and Oxford-102 flower dataset, and have demonstrated that our model is capable of synthesizing realistic images that match the given descriptions, while still maintain other features of original images.

****

Unified Deep Supervised Domain Adaptation and Generalization

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, Gianfranco Doretto; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5715-5725

This work addresses the problem of domain adaptation and generalization in a unified fashion. The main idea is to exploit the siamese architecture with the Contrastive Loss to address the domain shift and generalization problems. The framework is general, and can be used with any architecture. One of the main strengths of the approach is the "speed" of adaptation, which requires an extremely low number of labeled training samples from the target domain, even only one per category. The same architecture and loss function can be easily extended to domain generalization. We present state-of-the-art results for both of these applications.

****

Interpretable Transformations With Encoder-Decoder Networks

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, Gabriel J. Brostow; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5726-5735

Deep feature spaces have the capacity to encode complex transformations of their input data. However, understanding the relative feature-space relationship between two transformed encoded images is difficult. For instance, what is the relative feature space relationship between two rotated images? What is decoded when we interpolate in feature space? Ideally, we want to disentangle confounding factors, such as pose, appearance, and illumination, from object identity. Disentangling these is difficult because they interact in very nonlinear ways. We propose a simple method to construct a deep feature space, with explicitly disentangled representations of several known transformations. A person or algorithm can then manipulate the disentangled representation, for example, to re-render an image with explicit control over parameterized degrees of freedom. The feature space is constructed using a transforming encoder-decoder network with a custom feature transform layer, acting on the hidden representations. We demonstrate the advantages of explicit disentangling on a variety of datasets and transformations, and as an aid for traditional tasks, such as classification.

****

Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entro

py Minimization
Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, Heng Huang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5736-5745

In this paper, we propose a new clustering model, called DEeP Embedded RegularIzed ClusTering (DEPICT), which efficiently maps data into a discriminative embedding subspace and precisely predicts cluster assignments. DEPICT generally consists of a multinomial logistic regression function stacked on top of a multi-layer convolutional autoencoder. We define a clustering objective function using relative entropy (KL divergence) minimization, regularized by a prior for the frequency of cluster assignments. An alternating strategy is then derived to optimize the objective by updating parameters and estimating cluster assignments. Furthermore, we employ the reconstruction loss functions in our autoencoder, as a data-dependent regularization term, to prevent the deep embedding function from overfitting. In order to benefit from end-to-end optimization and eliminate the necessity for layer-wise pretraining, we introduce a joint learning framework to minimize the unified clustering and reconstruction loss functions together and train all network layers simultaneously. Experimental results indicate the superiority and faster running time of DEPICT in real-world clustering tasks, where no labeled data is available for hyper-parameter tuning.

****

Deep Scene Image Classification With the MFAFVNet
Yunsheng Li, Mandar Dixit, Nuno Vasconcelos; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5746-5754

The problem of transferring a deep convolutional network trained for object recognition to the task of scene image classification is considered. An embedded implementation of the recently proposed mixture of factor analyzers Fisher vector (MFA-FV) is proposed. This enables the design of a network architecture, the MFAFVNet, that can be trained in an end to end manner. The new architecture involves the design of an MFA-FV layer that implements a statistically correct version of the MFA-FV, through a combination of network computations and regularization. When compared to previous neural implementations of Fisher vectors, the MFAFVNet relies on a more powerful statistical model and a more accurate implementation. When compared to previous non-embedded models, the MFAFVNet relies on a state of the art model, which is now embedded into a CNN. This enables end to end training. Experiments show that the MFAFVNet has state of the art performance on scene classification.

****

Learning Bag-Of-Features Pooling for Deep Convolutional Neural Networks
Nikolaos Passalis, Anastasios Tefas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5755-5763

Convolutional Neural Networks (CNNs) are well established models capable of achieving state-of-the-art classification accuracy for various computer vision tasks. However, they are becoming increasingly larger, using millions of parameters, while they are restricted to handling images of fixed size. In this paper, a quantization-based approach, inspired from the well-known Bag-of-Features model, is proposed to overcome these limitations. The proposed approach, called Convolutional BoF (CBoF), uses RBF neurons to quantize the information extracted from the convolutional layers and it is able to natively classify images of various sizes as well as to significantly reduce the number of parameters in the network. In contrast to other global pooling operators and CNN compression techniques the proposed method utilizes a trainable pooling layer that it is end-to-end differentiable, allowing the network to be trained using regular back-propagation and to achieve greater distribution shift invariance than competitive methods. The ability of the proposed method to reduce the parameters of the network and increase the classification accuracy over other state-of-the-art techniques is demonstrated using three image datasets.

****

Adversarial Examples Detection in Deep Networks With Convolutional Filter Statistics

Xin Li, Fuxin Li; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5764-5772

Deep learning has greatly improved visual recognition in recent years. However, recent research has shown that there exist many adversarial examples that can negatively impact the performance of such an architecture. This paper focuses on detecting those adversarial examples by analyzing whether they come from the same distribution as the normal examples. Instead of directly training a deep neural network to detect adversarials, a much simpler approach was proposed based on statistics on outputs from convolutional layers. A cascade classifier was designed to efficiently detect adversarials. Furthermore, trained from one particular adversarial generating mechanism, the resulting classifier can successfully detect adversarials from a completely different mechanism as well. The resulting classifier is non-subdifferentiable, hence creates a difficulty for adversaries to attack by using the gradient of the classifier. After detecting adversarial examples, we show that many of them can be recovered by simply performing a small average filter on the image. Those findings should lead to more insights about the classification mechanisms in deep convolutional neural networks.

****

Joint Prediction of Activity Labels and Starting Times in Untrimmed Videos

Tahmida Mahmud, Mahmudul Hasan, Amit K. Roy-Chowdhury; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5773-5782

Most of the existing works on human activity analysis focus on recognition or early recognition of the activity labels from complete or partial observations. Predicting the labels of future unobserved activities where no frames of the predicted activities have been observed is a challenging problem, with important applications, which has not been explored much. Associated with the future label prediction problem is the problem of predicting the starting time of the next activity. In this work, we propose a system that is able to infer about the labels and the starting times of future activities. Activities are characterized by the previous activity sequence (which is observed), as well as the objects present in the scene during their occurrence. We propose a network similar to a hybrid Siamese network with three branches to jointly learn both the future label and the starting time. The first branch takes visual features from the objects present in the scene using a fully connected network, the second branch takes previous activity features using a LSTM network to model long-term sequential relationships and the third branch captures the last observed activity features to model the context of inter-activity time using another fully connected network. These concatenated features are used for both label and time prediction. Experiments on two challenging datasets demonstrate that our framework for joint prediction of activity label and starting time improves the performance of both, and outperforms the state-of-the-arts.

****

R-C3D: Region Convolutional 3D Network for Temporal Activity Detection

Huijuan Xu, Abir Das, Kate Saenko; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5783-5792

We address the problem of activity detection in continuous, untrimmed video streams. This is a difficult task that requires extracting meaningful spatio-temporal features to capture activities, accurately localizing the start and end times of each activity. We introduce a new model, Region Convolutional 3D Network (R-C3D), which encodes the video streams using a three-dimensional fully convolutional network, then generates candidate temporal regions containing activities, and finally classifies selected regions into specific activities. Computation is saved due to the sharing of convolutional features between the proposal and the classification pipelines. The entire model is trained end-to-end with jointly optimized localization and classification losses. R-C3D is faster than existing methods (569 frames per second on a single Titan X Maxwell GPU) and achieves state-of-the-art results on THUMOS'14. We further demonstrate that our model is a general activity detection framework that does not rely on assumptions about particular dataset properties by evaluating our approach on ActivityNet and Charades. Our code is available at http://ai.bu.edu/r-c3d/.

****

Temporal Context Network for Activity Localization in Videos

Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, Yan Qiu Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5793-5802

We present a Temporal Context Network (TCN) for precise temporal localization of human activities. Similar to the Faster-RCNN architecture, proposals are placed at equal intervals in a video which span multiple temporal scales. We propose a novel representation for ranking these proposals. Since pooling features only inside a segment is not sufficient to predict activity boundaries, we construct a representation which explicitly captures context around a proposal for ranking it. For each temporal segment inside a proposal, features are uniformly sampled at a pair of scales and are input to a temporal convolutional neural network for classification. After ranking proposals, non-maximum suppression is applied and classification is performed to obtain final detections. TCN outperforms state-of-the-art methods on the ActivityNet dataset and the THUMOS14 dataset.

****

Localizing Moments in Video With Natural Language

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, Bryan Russell; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5803-5812

We consider retrieving a specific temporal segment, or moment, from a video given a natural language text description. Methods designed to retrieve whole video clips with natural language determine what occurs in a video but not when. To address this issue, we propose the Moment Context Network (MCN) which effectively localizes natural language queries in videos by integrating local and global video features over time. A key obstacle to training our MCN model is that current video datasets do not include pairs of localized video segments and referring expressions, or text descriptions which uniquely identify a corresponding moment. Therefore, we collect the Distinct Describable Moments (DiDeMo) dataset which consists of over 10,000 unedited, personal videos in diverse visual settings with pairs of localized video segments and referring expressions. We demonstrate that MCN outperforms several baseline methods and believe that our initial results together with release of DiDeMo will inspire further research on localizing video moments with natural language.

****

TORNADO: A Spatio-Temporal Convolutional Regression Network for Video Action Proposal

Hongyuan Zhu, Romain Vial, Shijian Lu; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5813-5821

Given a video clip, action proposal aims to quickly generate a number of spatio-temporal tubes that enclose candidate human activities. Recently, the regression-based object detectors and long-term recurrent convolutional network (LRCN) have demonstrated superior performance in human action detection and recognition. However, the regression-based detectors performs inference without considering the temporal context among neighboring frames, and the LRCN using global visual percepts lacks the capability to capture local temporal dynamics. In this paper, we present a novel framework called TORNADO for human action proposal detection in un-trimmed video clips. Specifically, we propose a spatial-temporal convolutional network that combines the advantages of regression-based detector and LRCN by empowering Convolutional LSTM with regression capability. Our approach consists of a temporal convolutional regression network (T-CRN) and a spatial regression network (S-CRN) which are trained end-to-end on both RGB and OpticalFlow streams. They fuse appearance, motion and temporal contexts to regress the bounding boxes of candidate human actions simultaneously in 28 FPS. The action proposals are constructed by solving dynamic programming with peak trimming of the generated action boxes. Extensive experiments on the challenging UCF-101 and UCF-Sports datasets show that our method achieves superior performance as compared with the state-of-the-arts.

****

Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos

Rui Hou, Chen Chen, Mubarak Shah; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5822-5831

Deep learning has been demonstrated to achieve excellent results for image classification and object detection. However, the impact of deep learning on video analysis (e.g. action detection and recognition) has been limited due to complexity of video data and lack of annotations. Previous convolutional neural networks (CNN) based video action detection approaches usually consist of two major steps: frame-level action proposal detection and association of proposals across frames. Also, these methods employ two-stream CNN framework to handle spatial and temporal feature separately. In this paper, we propose an end-to-end deep network called Tube Convolutional Neural Network (T-CNN) for action detection in videos. The proposed architecture is a unified network that is able to recognize and localize action based on 3D convolution features. A video is first divided into equal length clips and for each clip a set of tube proposals are generated next based on 3D Convolutional Network (ConvNet) features. Finally, the tube proposals of different clips are linked together employing network flow and spatio-temporal action detection is performed using these linked video proposals. Extensive experiments on several video datasets demonstrate the superior performance of T-CNN for classifying and localizing actions in both trimmed and untrimmed videos compared to state-of-the-arts.

****

Learning Action Recognition Model From Depth and Skeleton Videos

Hossein Rahmani, Mohammed Bennamoun; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5832-5841

Depth sensors open up possibilities of dealing with the human action recognition problem by providing 3D human skeleton data and depth images of the scene. Analysis of human actions based on 3D skeleton data has become popular recently, due to its robustness and view-invariant representation. However, the skeleton alone is insufficient to distinguish actions which involve human-object interactions. In this paper, we propose a deep model which efficiently models human-object interactions and intra-class variations under viewpoint changes. First, a human body-part model is introduced to transfer the depth appearances of body-parts to a shared view-invariant space. Second, an end-to-end learning framework is proposed which is able to effectively combine the view-invariant body-part representation from skeletal and depth images, and learn the relations between the human body-parts and the environmental objects, the interactions between different human body-parts, and the temporal structure of human actions. We have evaluated the performance of our proposed model against 15 existing techniques on two large benchmark human action recognition datasets including NTU RGB+D and UWA3DII. The Experimental results show that our technique provides a significant improvement over state-of-the-art methods.

****

The "Something Something" Video Database for Learning and Evaluating Visual Common Sense

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, Roland Memisevic; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5842-5850

Neural networks trained on datasets such as ImageNet have led to major advances in visual object classification. One obstacle that prevents networks from reasoning more deeply about complex scenes and situations, and from integrating visual knowledge with natural language, like humans do, is their lack of common sense knowledge about the physical world. Videos, unlike still images, contain a wealth of detailed information about the physical world. However, most labelled video datasets represent high-level concepts rather than detailed physical aspects about actions and scenes. In this work, we describe our ongoing collection of the "something-something" database of video prediction tasks whose solutions require a common sense understanding of the depicted situation. The database currently contains more than 100,000 videos across 174 classes, which are defined as caption-templates. We also describe the challenges in crowd-sourcing this data at scale.

****

GPLAC: Generalizing Vision-Based Robotic Skills Using Weakly Labeled Images

Avi Singh, Larry Yang, Sergey Levine; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5851-5860

We tackle the problem of learning robotic sensorimotor control policies that can generalize to visually diverse and unseen environments. Achieving broad generalization typically requires large datasets, which are difficult to obtain for task-specific interactive processes such as reinforcement learning or learning from demonstration. However, much of the visual diversity in the world can be captured through passively collected datasets of images or videos. In our method, which we refer to as GPLAC (Generalized Policy Learning with Attentional Classifier), we use both interaction data and weakly labeled image data to augment the generalization capacity of sensorimotor policies. Our method combines multitask learning on action selection and an auxiliary binary classification objective, together with a convolutional neural network architecture that uses an attentional mechanism to avoid distractors. We show that pairing interaction data from just a single environment with a diverse dataset of weakly labeled data results in greatly improved generalization to unseen environments, and show that this generalization depends on both the auxiliary objective and the attentional architecture that we propose. We demonstrate our results in both simulation and on a real robotic manipulator, and demonstrate substantial improvement over standard convolutional architectures and domain adaptation methods.

****

Semi-Global Weighted Least Squares in Image Filtering

Wei Liu, Xiaogang Chen, Chuanhua Shen, Zhi Liu, Jie Yang; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5861-5869

Solving the global method of Weighted Least Squares (WLS) model in image filtering is both time- and memory-consuming. In this paper, we present an alternative approximation in a time- and memory- efficient manner which is denoted as Semi-Global Weighed Least Squares (SG-WLS). Instead of solving a large linear system, we propose to iteratively solve a sequence of subsystems which are one-dimensional WLS models. Although each subsystem is one-dimensional, it can take two-dimensional neighborhood information into account due to the proposed special neighborhood construction. We show such a desirable property makes our SG-WLS achieve close performance to the original two-dimensional WLS model but with much less time and memory cost. While previous related methods mainly focus on the 4-connected/8-connected neighborhood system, our SG-WLS can handle a more general and larger neighborhood system thanks to the proposed fast solution. We show such a generalization can achieve better performance than the 4-connected/8-connected neighborhood system in some applications. Our SG-WLS is ~20 times faster than the WLS model. For an image of MxN, the memory cost of SG-WLS is at most at the magnitude of max\ 1 / M, 1 / N\ of that of the WLS model. We show the effectiveness and efficiency of our SG-WLS in a range of applications.

****

Scale Recovery for Monocular Visual Odometry Using Depth Estimated With Deep Convolutional Neural Fields

Xiaochuan Yin, Xiangwei Wang, Xiaoguo Du, Qijun Chen; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5870-5878

Scale recovery is one of the central problems for monocular visual odometry. Normally, road plane and camera height are specified as reference to recover the scale. The performances of these methods depend on the plane recognition and height measurement of camera. In this work, we propose a novel method to recover the scale by incorporating the depths estimated from images using deep convolutional neural fields. Our method considers the whole environmental structure as reference rather than a specified plane. The accuracy of depth estimation contributes to the scale recovery. We improve the performance of depth estimation by considering two consecutive frames and egomotion of camera into our networks. The depth refinement and scale recovery are obtained iteratively. In this way, our method can eliminate the scale drift and improve the depth estimation simultaneously. The effectiveness of our method is verified on the KITTI dataset for both visual odometry and depth estimation tasks.

****

Deep Adaptive Image Clustering

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5879-5887

Image clustering is a crucial but challenging task in machine learning and computer vision. Existing methods often ignore the combination between feature learning and clustering. To tackle this problem, we propose Deep Adaptive Clustering (DAC) that recasts the clustering problem into a binary pairwise-classification framework to judge whether pairs of images belong to the same clusters. In DAC, the similarities are calculated as the cosine distance between label features of images which are generated by a deep convolutional network (ConvNet). By introducing a constraint into DAC, the learned label features tend to be one-hot vectors that can be utilized for clustering images. The main challenge is that the ground-truth similarities are unknown in image clustering. We handle this issue by presenting an alternating iterative Adaptive Learning algorithm where each iteration alternately selects labeled samples and trains the ConvNet. Conclusively, images are automatically clustered based on the label features. Experimental results show that DAC achieves state-of-the-art performance on five popular datasets, e.g., yielding 97.75% clustering accuracy on MNIST, 52.18% on CIFAR-10 and 46.99% on STL-10.

****

One Network to Solve Them All -- Solving Linear Inverse Problems Using Deep Projection Models

J. H. Rick Chang, Chun-Liang Li, Barnabas Poczos, B. V. K. Vijaya Kumar, Aswin C. Sankaranarayanan; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5888-5897

While deep learning methods have achieved state-of-the-art performance in many challenging inverse problems like image inpainting and super-resolution, they invariably involve problem-specific training of the networks. Under this approach, each inverse problem requires its own dedicated network. In scenarios where we need to solve a wide variety of problems, e.g., on a mobile camera, it is inefficient and expensive to use these problem-specific networks. On the other hand, traditional methods using analytic signal priors can be used to solve any linear inverse problem; this often comes with a performance that is worse than learning-based methods. In this work, we provide a middle ground between the two kinds of methods -- we propose a general framework to train a single deep neural network that solves arbitrary linear inverse problems. We achieve this by training a network that acts as a quasi-projection operator for the set of natural images and show that any linear inverse problem involving natural images can be solved using iterative methods. We empirically show that the proposed framework demonstrates superior performance over traditional methods using wavelet sparsity prior while achieving performance comparable to specially-trained networks on tasks including compressive sensing and pixel-wise inpainting.

****

Representation Learning by Learning to Count

Mehdi Noroozi, Hamed Pirsiavash, Paolo Favaro; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5898-5906

We introduce a novel method for representation learning that uses an artificial supervision signal based on counting visual primitives. This supervision signal is obtained from an equivariance relation, which does not require any manual annotation. We relate transformations of images to transformations of the representations. More specifically, we look for the representation that satisfies such relation rather than the transformations that match a given representation. In this paper, we use two image transformations in the context of counting: scaling and tiling. The first transformation exploits the fact that the number of visual p rimitives should be invariant to scale. The second transformation allows us to equate the total number of visual primitives in each tile to that in the whole image. These two transformations are combined in one constraint and used to train a neural network with a contrastive loss. The proposed task produces representations that perform on par or exceed the state of the art in transfer learning benchmarks.

****

StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris N. Metaxas; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5907-5915

Synthesizing high-quality images from text descriptions is a challenging problem in computer vision and has many practical applications. Samples generated by existing text-to-image approaches can roughly reflect the meaning of the given descriptions, but they fail to contain necessary details and vivid object parts. In this paper, we propose Stacked Generative Adversarial Networks (StackGAN) to generate 256x256 photo-realistic images conditioned on text descriptions. We decompose the hard problem into more manageable sub-problems through a sketch-refinement process. The Stage-I GAN sketches the primitive shape and colors of the object based on the given text description, yielding Stage-I low-resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details. It is able to rectify defects in Stage-I results and add compelling details with the refinement process. To improve the diversity of the synthesized images and stabilize the training of the conditional-GAN, we introduce a novel Conditioning Augmentation technique that encourages smoothness in the latent conditioning manifold. Extensive experiments and comparisons with state-of-the-arts on benchmark datasets demonstrate that the proposed method achieves significant improvements on generating photo-realistic images conditioned on text descriptions.

****

Unsupervised Learning of Object Landmarks by Factorized Spatial Embeddings

James Thewlis, Hakan Bilen, Andrea Vedaldi; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5916-5925

Automatically learning the structure of object categories remains an important open problem in computer vision. We propose a novel unsupervised approach that can discover and learn to detect landmarks in object categories, thus characterizing their structure. Our approach is based on factorizing image deformations, as induced by a viewpoint change or an object articulation, by learning a deep neural network that detects landmarks compatible with such visual effects. We show that, by requiring the same neural network to be applicable to different object instances, our method naturally induces meaningful correspondences between different object instances in a category. We assess the method qualitatively on a variety of object types, natural an man-made. We also show that our unsupervised landmarks are highly predictive of manually-annotated landmarks in faces benchmark datasets, and can be used to regress those with a high degree of accuracy.

****