Fast and Accurate Image Matching with Cascade Hashing for 3D Reconstruction

Jian Cheng, Cong Leng, Jiaxiang Wu, Hainan Cui, Hanqing Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1-8

Image matching is one of the most challenging stages in 3D reconstruction, which usually occupies half of computational cost and inaccurate matching may lead to failure of reconstruction. Therefore, fast and accurate image matching is very crucial for 3D reconstruction. In this paper, we proposed a Cascade Hashing strategy to speed up the image matching. In order to accelerate the image matching, the proposed Cascade Hashing method is designed to be three-layer structure: hashing lookup, hashing remapping, and hashing ranking. Each layer adopts different measures and filtering strategies, which is demonstrated to be less sensitive to noise. Extensive experiments show that image matching can be accelerated by our approach in hundreds times than brute force matching, even achieves ten times or more than Kd-tree based matching while retaining comparable accuracy.

****

Predicting Matchability

Wilfried Hartmann, Michal Havlena, Konrad Schindler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 9-16

The initial steps of many computer vision algorithms are interest point extraction and matching. In larger image sets the pairwise matching of interest point descriptors between images is an important bottleneck. For each descriptor in one image the (approximate) nearest neighbor in the other one has to be found and checked against the second-nearest neighbor to ensure the correspondence is unambiguous. Here, we asked the question how to best decimate the list of interest points without losing matches, i.e. we aim to speed up matching by filtering out, in advance, those points which would not survive the matching stage. It turns out that the best filtering criterion is not the response of the interest point detector, which in fact is not surprising: the goal of detection are repeatable and well-localized points, whereas the objective of the selection are points whose descriptors can be matched successfully. We show that one can in fact learn to predict which descriptors are matchable, and thus reduce the number of interest points significantly without losing too many matches. We show that this strategy, as simple as it is, greatly improves the matching success with the same number of points per image. Moreover, we embed the prediction in a state-of-the-art Structure-from-Motion pipeline and demonstrate that it also outperforms other selection methods at system level.

****

Trinocular Geometry Revisited

Jean Ponce, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 17-24

When do the visual rays associated with triplets of point correspondences converge, that is, intersect in a common point? Classical models of trinocular geometry based on the fundamental matrices and trifocal tensor associated with the corresponding cameras only provide partial answers to this fundamental question, in large part because of underlying, but seldom explicit, general configuration assumptions. This paper uses elementary tools from projective line geometry to provide necessary and sufficient geometric and analytical conditions for convergence in terms of transversals to triplets of visual rays, without any such assumptions. In turn, this yields a novel and simple minimal parameterization of trinocular geometry for cameras with non-collinear or collinear pinholes.

****

Critical Configurations For Radial Distortion Self-Calibration

Changchang Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 25-32

In this paper, we study the configurations of motion and structure that lead to inherent ambiguities in radial distortion estimation (or 3D reconstruction with unknown radial distortions). By analyzing the motion field of radially distorted images, we solve for critical surface pairs that can lead to the same motion field under different radial distortions and possibly different camera motions. We study the properties of the discovered critical configurations and discuss the practically important configurations that often occur in real applications. We demonstrate the impact of the radial distortion ambiguity on multi-view reconstruction with synthetic experiments and real experiments.

****

Minimal Solvers for Relative Pose with a Single Unknown Radial Distortion
Yubin Kuang, Jan E. Solem, Fredrik Kahl, Kalle Astrom; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 33-40

In this paper, we study the problems of estimating relative pose between two cameras in the presence of radial distortion. Specifically, we consider minimal problems where one of the cameras has no or known radial distortion. There are three useful cases for this setup with a single unknown distortion: (i) fundamental matrix estimation where the two cameras are uncalibrated, (ii) essential matrix estimation for a partially calibrated camera pair, (iii) essential matrix estimation for one calibrated camera and one camera with unknown focal length. We study the parameterization of these three problems and derive fast polynomial solvers based on Grobner basis methods. We demonstrate the numerical stability of the solvers on synthetic data. The minimal solvers have also been applied to real imagery with convincing results

****

Reconstructing PASCAL VOC
Sara Vicente, Joao Carreira, Lourdes Agapito, Jorge Batista; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 41-48

We address the problem of populating object category detection datasets with dense, per-object 3D reconstructions, bootstrapped from class labels, ground truth figure-ground segmentations and a small set of keypoint annotations. Our proposed algorithm first estimates camera viewpoint using rigid structure-from-motion, then reconstructs object shapes by optimizing over visual hull proposals guided by loose within-class shape similarity assumptions. The visual hull sampling process attempts to intersect an object's projection cone with the cones of minimal subsets of other similar objects among those pictured from certain vantage points. We show that our method is able to produce convincing per-object 3D reconstructions on one of the most challenging existing object-category detection datasets, PASCAL VOC. Our results may re-stimulate once popular geometry-oriented model-based recognition approaches.

****

Spectral Graph Reduction for Efficient Image and Streaming Video Segmentation
Fabio Galasso, Margret Keuper, Thomas Brox, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 49-56

Computational and memory costs restrict spectral techniques to rather small graphs, which is a serious limitation especially in video segmentation. In this paper, we propose the use of a reduced graph based on superpixels. In contrast to previous work, the reduced graph is reweighted such that the resulting segmentation is equivalent, under certain assumptions, to that of the full graph. We consider equivalence in terms of the normalized cut and of its spectral clustering relaxation. The proposed method reduces runtime and memory consumption and yields on par results in image and video segmentation. Further, it enables an efficient data representation and update for a new streaming video segmentation approach that also achieves state-of-the-art performance.

****

Weakly Supervised Multiclass Video Segmentation
Xiao Liu, Dacheng Tao, Mingli Song, Ying Ruan, Chun Chen, Jiajun Bu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 57-64

The desire of enabling computers to learn semantic concepts from large quantities of Internet videos has motivated increasing interests on semantic video understanding, while video segmentation is important yet challenging for understanding videos. The main difficulty of video segmentation arises from the burden of labeling training samples, making the problem largely unsolved. In this paper, we present a novel nearest neighbor-based label transfer scheme for weakly supervise d video segmentation. Whereas previous weakly supervised video segmentation methods have been limited to the two-class case, our proposed scheme focuses on more challenging multiclass video segmentation, which finds a semantically meaningful label for every pixel in a video. Our scheme enjoys several favorable properties when compared with conventional methods. First, a weakly supervised hashing procedure is carried out to handle both metric and semantic similarity. Second, the proposed nearest neighbor-based label transfer algorithm effectively avoids overfitting caused by weakly supervised data. Third, a multi-video graph model is built to encourage smoothness between regions that are spatiotemporally adjacent and similar in appearance. We demonstrate the effectiveness of the proposed scheme by comparing it with several other state-of-the-art weakly supervised segmentation methods on one new Wild8 dataset and two other publicly available datasets.

****

Video Motion Segmentation Using New Adaptive Manifold Denoising Model

Dijun Luo, Heng Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 65-72

Video motion segmentation techniques automatically segment and track objects and regions from videos or image sequences as a primary processing step for many computer vision applications. We propose a novel motion segmentation approach for both rigid and non-rigid objects using adaptive manifold denoising. We first introduce an adaptive kernel space in which two feature trajectories are mapped into the same point if they belong to the same rigid object. After that, we employ an embedded manifold denoising approach with the adaptive kernel to segment the motion of rigid and non-rigid objects. The major observation is that the non-rigid objects often lie on a smooth manifold with deviations which can be removed by manifold denoising. We also show that performing manifold denoising on the kernel space is equivalent to doing so on its range space, which theoretically justifies the embedded manifold denoising on the adaptive kernel space. Experimental results indicate that our algorithm, named Adaptive Manifold Denoising (AMD), is suitable for both rigid and non-rigid motion segmentation. Our algorithm works well in many cases where several state-of-the-art algorithms fail.

****

Cut, Glue & Cut: A Fast, Approximate Solver for Multicut Partitioning

Thorsten Beier, Thorben Kroeger, Jorg H. Kappes, Ullrich Kothe, Fred A. Hamprecht; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 73-80

Recently, unsupervised image segmentation has become increasingly popular. Starting from a superpixel segmentation, an edge-weighted region adjacency graph is constructed. Amongst all segmentations of the graph, the one which best conforms to the given image evidence, as measured by the sum of cut edge weights, is chosen. Since this problem is NP-hard, we propose a new approximate solver based on the move-making paradigm: first, the graph is recursively partitioned into small regions (cut phase). Then, for any two adjacent regions, we consider alternative cuts of these two regions defining possible moves (glue & cut phase). For planar problems, the optimal move can be found, whereas for non-planar problems, efficient approximations exist. We evaluate our algorithm on published and new benchmark datasets, which we make available here. The proposed algorithm finds segmentations that, as measured by a loss function, are as close to the ground-truth as the global optimum found by exact solvers. It does so significantly faster then existing approximate methods, which is important for large-scale problems.

****

Neural Decision Forests for Semantic Image Labelling

Samuel Rota Bulo, Peter Kontschieder; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 81-88

In this work we present Neural Decision Forests, a novel approach to jointly tackle data representation- and discriminative learning within randomized decision trees. Recent advances of deep learning architectures demonstrate the power of embedding representation learning within the classifier â■■ An idea that is intuitively supported by the hierarchical nature of the decision forest model where t he input space is typically left unchanged during training and testing. We bridge this gap by introducing randomized Multi- Layer Perceptrons (rMLP) as new split nodes which are capable of learning non-linear, data-specific representations and taking advantage of them by finding optimal predictions for the emerging child nodes. To prevent overfitting, we i) randomly select the image data fed to the input layer, ii) automatically adapt the rMLP topology to meet the complexity of the data arriving at the node and iii) introduce an l1-norm based regularization that additionally sparsifies the network. The key findings in our experiments on three different semantic image labelling datasets are consistently improved results and significantly compressed trees compared to conventional classification trees.

****

Pulling Things out of Perspective

Lubor Ladicky, Jianbo Shi, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 89-96

The limitations of current state-of-the-art methods for single-view depth estimation and semantic segmentations are closely tied to the property of perspective geometry, that the perceived size of the objects scales inversely with the distance. In this paper, we show that we can use this property to reduce the learning of a pixel-wise depth classifier to a much simpler classifier predicting only the likelihood of a pixel being at an arbitrarily fixed canonical depth. The likelihoods for any other depths can be obtained by applying the same classifier after appropriate image manipulations. Such transformation of the problem to the canonical depth removes the training data bias towards certain depths and the effect of perspective. The approach can be straight-forwardly generalized to multiple semantic classes, improving both depth estimation and semantic segmentation performance by directly targeting the weaknesses of independent approaches. Conditioning the semantic label on the depth provides a way to align the data to their physical scale, allowing to learn a more discriminative classifier. Conditioning depth on the semantic class helps the classifier to distinguish between ambiguities of the otherwise ill-posed problem. We tested our algorithm on the KITTI road scene dataset and NYU2 indoor dataset and obtained obtained results that significantly outperform current state-of-the-art in both single-view depth and semantic segmentation domain.

****

Event Detection using Multi-Level Relevance Labels and Multiple Features

Zhongwen Xu, Ivor W. Tsang, Yi Yang, Zhigang Ma, Alexander G. Hauptmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 97-104

We address the challenging problem of utilizing related exemplars for complex event detection while multiple features are available. Related exemplars share certain positive elements of the event, but have no uniform pattern due to the huge variance of relevance levels among different related exemplars. None of the existing multiple feature fusion methods can deal with the related exemplars. In this paper, we propose an algorithm which adaptively utilizes the related exemplars by cross-feature learning. Ordinal labels are used to represent the multiple relevance levels of the related videos. Label candidates of related exemplars are generated by exploring the possible relevance levels of each related exemplar via a cross-feature voting strategy. Maximum margin criterion is then applied in our framework to discriminate the positive and negative exemplars, as well as the related exemplars from different relevance levels. We test our algorithm using the large scale TRECVID 2011 dataset and it gains promising performance.

****

Full-Angle Quaternions for Robustly Matching Vectors of 3D Rotations

Stephan Liwicki, Minh-Tri Pham, Stefanos Zafeiriou, Maja Pantic, Bjorn Stenger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 105-112

In this paper we introduce a new distance for robustly matching vectors of 3D rotations. A special representation of 3D rotations, which we coin full-angle quaternion (FAQ), allows us to express this distance as Euclidean. We apply the dist ance to the problems of 3D shape recognition from point clouds and 2D object tracking in color video. For the former, we introduce a hashing scheme for scale and translation which outperforms the previous state-of-the-art approach on a public dataset. For the latter, we incorporate online subspace learning with the proposed FAQ representation to highlight the benefits of the new representation.

****

Human vs. Computer in Scene and Object Recognition

Ali Borji, Laurent Itti; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 113-120

Several decades of research in computer and primate vision have resulted in many models (some specialized for one problem, others more general) and invaluable experimental data. Here, to help focus research efforts onto the hardest unsolved problems, and bridge computer and human vision, we define a battery of 5 tests that measure the gap between human and machine performances in several dimensions (generalization across scene categories, generalization from images to edge maps and line drawings, invariance to rotation and scaling, local/global information with jumbled images, and object recognition performance). We measure model accuracy and the correlation between model and human error patterns. Experimenting over 7 datasets, where human data is available, and gauging 14 well-established models, we find that none fully resembles humans in all aspects, and we learn from each test which models and features are more promising in approaching humans in the tested dimension. Across all tests, we find that models based on local edge histograms consistently resemble humans more, while several scene statistics or "gist" models do perform well with both scenes and objects. While computer vision has long been inspired by human vision, we believe systematic efforts, such as this, will help better identify shortcomings of models and find new paths forward.

****

Semi-supervised Spectral Clustering for Image Set Classification

Arif Mahmood, Ajmal Mian, Robyn Owens; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 121-128

We present an image set classification algorithm based on unsupervised clustering of labeled training and unlabeled test data where labels are only used in the stopping criterion. The probability distribution of each class over the set of clusters is used to define a true set based similarity measure. To this end, we propose an iterative sparse spectral clustering algorithm. In each iteration, a proximity matrix is efficiently recomputed to better represent the local subspace structure. Initial clusters capture the global data structure and finer clusters at the later stages capture the subtle class differences not visible at the global scale. Image sets are compactly represented with multiple Grassmannian manifolds which are subsequently embedded in Euclidean space with the proposed spectral clustering algorithm. We also propose an efficient eigenvector solver which not only reduces the computational cost of spectral clustering by many folds but also improves the clustering quality and final classification results. Experiments on five standard datasets and comparison with seven existing techniques show the efficacy of our algorithm.

****

Look at the Driver, Look at the Road: No Distraction! No Accident!

Mahdi Rezaei, Reinhard Klette; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 129-136

The paper proposes an advanced driver-assistance system that correlates the driver's head pose to road hazards by analyzing both simultaneously. In particular, we aim at the prevention of rear-end crashes due to driver fatigue or distraction. We contribute by three novel ideas: Asymmetric appearance-modeling, 2D to 3D pose estimation enhanced by the introduced Fermat-point transform, and adaptation of Global Haar (GHaar) classifiers for vehicle detection under challenging lighting conditions. The system defines the driver's direction of attention (in 6 degrees of freedom), yawning and head-nodding detection, as well as vehicle detection, and distance estimation. Having both road and driver's behaviour information, and implementing a fuzzy fusion system, we develop an integrated framework t o cover all of the above subjects. We provide real-time performance analysis for real-world driving scenarios.

****

Measuring Distance Between Unordered Sets of Different Sizes

Andrew Gardner, Jinko Kanno, Christian A. Duncan, Rastko Selmic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 137-143

We present a distance metric based upon the notion of minimum-cost injective mappings between sets. Our function satisfies metric properties as long as the cost of the minimum mappings is derived from a semimetric, for which the triangle inequality is not necessarily satisfied. We show that the Jaccard distance (alternatively biotope, Tanimoto, or Marczewski-Steinhaus distance) may be considered the special case for finite sets where costs are derived from the discrete metric. Extensions that allow premetrics (not necessarily symmetric), multisets (generalized to include probability distributions), and asymmetric mappings are given that expand the versatility of the metric without sacrificing metric properties. The function has potential applications in pattern recognition, machine learning, and information retrieval.

****

Learning Mid-level Filters for Person Re-identification

Rui Zhao, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 144-151

In this paper, we propose a novel approach of learning mid-level filters from automatically discovered patch clusters for person re-identification. It is well motivated by our study on what are good filters for person re-identification. Our mid-level filters are discriminatively learned for identifying specific visual patterns and distinguishing persons, and have good cross-view invariance. First, local patches are qualitatively measured and classified with their discriminative power. Discriminative and representative patches are collected for filter learning. Second, patch clusters with coherent appearance are obtained by pruning hierarchical clustering trees, and a simple but effective cross-view training strategy is proposed to learn filters that are view-invariant and discriminative. Third, filter responses are integrated with patch matching scores in RankSVM training. The effectiveness of our approach is validated on the VIPeR dataset and the CUHK01 dataset. The learned mid-level features are complementary to existing handcrafted low-level features, and improve the best Rank-1 matching rate on the VIPeR dataset by 14%.

****

DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification

Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 152-159

Person re-identification is to match pedestrian images from disjoint camera views detected by pedestrian detectors. Challenges are presented in the form of complex variations of lightings, poses, viewpoints, blurring effects, image resolutions, camera settings, occlusions and background clutter across camera views. In addition, misalignment introduced by the pedestrian detector will affect most existing person re-identification methods that use manually cropped pedestrian images and assume perfect detection. In this paper, we propose a novel filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter. All the key components are jointly optimized to maximize the strength of each component when cooperating with others. In contrast to existing works that use handcrafted features, our method automatically learns features optimal for the re-identification task from data. The learned filter pairs encode photometric transforms. Its deep architecture makes it possible to model a mixture of complex photometric and geometric transforms. We build the largest benchmark re-id dataset with 13,164 images of 1,360 pedestrians. Unlike existing datasets, which only provide manually cropped pedestrian images, our dataset provides automatically detected bounding boxes for evaluation close to practical applications. Our neural network significantly outperforms state-of-the-art methods on this dataset.

****

Lacunarity Analysis on Image Patterns for Texture Classification

Yuhui Quan, Yong Xu, Yuping Sun, Yu Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 160-167

Based on the concept of lacunarity in fractal geometry, we developed a statistical approach to texture description, which yields highly discriminative feature with strong robustness to a wide range of transformations, including photometric changes and geometric changes. The texture feature is constructed by concatenating the lacunarity-related parameters estimated from the multi-scale local binary patterns of image. Benefiting from the ability of lacunarity analysis to distinguish spatial patterns, our method is able to characterize the spatial distribution of local image structures from multiple scales. The proposed feature was applied to texture classification and has demonstrated excellent performance in comparison with several state-of-the-art approaches on four benchmark datasets.

****

Segmentation-aware Deformable Part Models

Eduard Trulls, Stavros Tsogkas, Iasonas Kokkinos, Alberto Sanfeliu, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 168-175

In this work we propose a technique to combine bottom-up segmentation, coming in the form of SLIC superpixels, with sliding window detectors, such as Deformable Part Models (DPMs). The merit of our approach lies in "cleaning up" the low-level HOG features by exploiting the spatial support of SLIC superpixels; this can be understood as using segmentation to split the feature variation into object-specific and background changes. Rather than committing to a single segmentation we use a large pool of SLIC superpixels and combine them in a scale-, position- and object-dependent manner to build soft segmentation masks. The segmentation masks can be computed fast enough to repeat this process over every candidate window, during training and detection, for both the root and part filters of DPMs. We use these masks to construct enhanced, background-invariant features to train DPMs. We test our approach on the PASCAL VOC 2007, outperforming the standard DPM in 17 out of 20 classes, yielding an average increase of 1.7% AP. Additionally, we demonstrate the robustness of this approach, extending it to dense SIFT descriptors for large displacement optical flow.

****

From Categories to Individuals in Real Time -- A Unified Boosting Approach

David Hall, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 176-183

A method for online, real-time learning of individual-object detectors is presented. Starting with a pre-trained boosted category detector, an individual-object detector is trained with near-zero computational cost. The individual detector is obtained by using the same feature cascade as the category detector along with elementary manipulations of the thresholds of the weak classifiers. This is ideal for online operation on a video stream or for interactive learning. Applications addressed by this technique are reidentification and individual tracking.
Experiments on four challenging pedestrian and face datasets indicate that it is indeed possible to learn identity classifiers in real-time; besides being faster-trained, our classifier has better detection rates than previous methods on two of the datasets.

****

NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization

Mahdi M. Kalayeh, Haroon Idrees, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 184-191

The real world image databases such as Flickr are characterized by continuous addition of new images. The recent approaches for image annotation, i.e. the problem of assigning tags to images, have two major drawbacks. First, either models are learned using the entire training data, or to handle the issue of dataset imbalance, tag-specific discriminative models are trained. Such models become obsolete and require relearning when new images and tags are added to database. Secon

d, the task of feature-fusion is typically dealt using ad-hoc approaches. In this paper, we present a weighted extension of Multi-view Non-negative Matrix Factorization (NMF) to address the aforementioned drawbacks. The key idea is to learn query-specific generative model on the features of nearest-neighbors and tags using the proposed NMF-KNN approach which imposes consensus constraint on the coefficient matrices across different features. This results in coefficient vectors across features to be consistent and, thus, naturally solves the problem of feature fusion, while the weight matrices introduced in the proposed formulation alleviate the issue of dataset imbalance. Furthermore, our approach, being query-specific, is unaffected by addition of images and tags in a database. We tested our method on two datasets used for evaluation of image annotation and obtained competitive results.
****

Fine-Grained Visual Comparisons with Local Learning
Aron Yu, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 192-199

Given two images, we want to predict which exhibits a particular visual attribute more than the other---even when the two images are quite similar. Existing relative attribute methods rely on global ranking functions; yet rarely will the visual cues relevant to a comparison be constant for all data, nor will humans' perception of the attribute necessarily permit a global ordering. To address these issues, we propose a local learning approach for fine-grained visual comparisons. Given a novel pair of images, we learn a local ranking model on the fly, using only analogous training comparisons. We show how to identify these analogous pairs using learned metrics. With results on three challenging datasets -- including a large newly curated dataset for fine-grained comparisons -- our method outperforms state-of-the-art methods for relative attribute prediction.
****

Inferring Analogous Attributes
Chao-Yeh Chen, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 200-207

The appearance of an attribute can vary considerably from class to class (e.g., a "fluffy" dog vs. a "fluffy" towel), making standard class-independent attribute models break down. Yet, training object-specific models for each attribute can be impractical, and defeats the purpose of using attributes to bridge category boundaries. We propose a novel form of transfer learning that addresses this dilemma. We develop a tensor factorization approach which, given a sparse set of class-specific attribute classifiers, can infer new ones for object-attribute pairs unobserved during training. For example, even though the system has no labeled images of striped dogs, it can use its knowledge of other attributes and objects to tailor "stripedness" to the dog category. With two large-scale datasets, we demonstrate both the need for category-sensitive attributes as well as our method's successful transfer. Our inferred attribute classifiers perform similarly well to those trained with the luxury of labeled class-specific instances, and much better than those restricted to traditional modes of transfer.
****

Beyond Comparing Image Pairs: Setwise Active Learning for Relative Attributes
Lucy Liang, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 208-215

It is useful to automatically compare images based on their visual properties---to predict which image is brighter, more feminine, more blurry, etc. However, comparative models are inherently more costly to train than their classification counterparts. Manually labeling all pairwise comparisons is intractable, so which pairs should a human supervisor compare? We explore active learning strategies for training relative attribute ranking functions, with the goal of requesting human comparisons only where they are most informative. We introduce a novel criterion that requests a partial ordering for a set of examples that minimizes the total rank margin in attribute space, subject to a visual diversity constraint. The setwise criterion helps amortize effort by identifying mutually informative comparisons, and the diversity requirement safeguards against requests a hu man viewer will find ambiguous. We develop an efficient strategy to search for sets that meet this criterion. On three challenging datasets and experiments with "live" online annotators, the proposed method outperforms both traditional passive learning as well as existing active rank learning methods.

****

Visual Persuasion: Inferring Communicative Intents of Images
Jungseock Joo, Weixin Li, Francis F. Steen, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 216-223

In this paper we introduce the novel problem of understanding visual persuasion. Modern mass media make extensive use of images to persuade people to make commercial and political decisions. These effects and techniques are widely studied in the social sciences, but behavioral studies do not scale to massive datasets. Computer vision has made great strides in building syntactical representations of images, such as detection and identification of objects. However, the pervasive use of images for communicative purposes has been largely ignored. We extend the significant advances in syntactic analysis in computer vision to the higher-level challenge of understanding the underlying communicative intent implied in images. We begin by identifying nine dimensions of persuasive intent latent in images of politicians, such as "socially dominant," "energetic," and "trustworthy," and propose a hierarchical model that builds on the layer of syntactical attributes, such as "smile" and "waving hand," to predict the intents presented in the images. To facilitate progress, we introduce a new dataset of 1,124 images of politicians labeled with ground-truth intents in the form of rankings. This study demonstrates that a systematic focus on visual persuasion opens up the field of computer vision to a new class of investigations around mediated images, intersecting with media analysis, psychology, and political communication.

****

Histograms of Pattern Sets for Image Classification and Object Recognition
Winn Voravuthikunchai, Bruno Cremilleux, Frederic Jurie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 224-231

This paper introduces a novel image representation capturing feature dependencies through the mining of meaningful combinations of visual   features. This representation leads to a compact and discriminative encoding of images that can be used for image classification, object detection or object recognition. The method relies on (i) multiple random projections of the input space followed by local binarization of projected histograms encoded as sets of items, and (ii) the representation of images as Histograms of Pattern Sets (HoPS). The approach is validated on four publicly available datasets (Daimler Pedestrian, Oxford Flowers, KTH Texture and PASCAL VOC2007), allowing comparisons with many recent approaches. The proposed image representation reaches state-of-the-art performance on each one of these datasets.

****

Incorporating Scene Context and Object Layout into Appearance Modeling
Hamid Izadinia, Fereshteh Sadeghi, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 232-239

A scene category imposes tight distributions over the kind of objects that might appear in the scene, the appearance of those objects and their layout. In this paper, we propose a method to learn scene structures that can encode three main interlacing components of a scene: the scene category, the context-specific appearance of objects, and their layout. Our experimental evaluations show that our learned scene structures outperform state-of-the-art method of Deformable Part Models in detecting objects in a scene. Our scene structure provides a level of scene understanding that is amenable to deep visual inferences. The scene structures can also generate features that can later be used for scene categorization. Using these features, we also show promising results on scene categorization.

****

Co-Segmentation of Textured 3D Shapes with Sparse Annotations
Mehmet Ersin Yumer, Won Chun, Ameesh Makadia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 240-247

We present a novel co-segmentation method for textured 3D shapes. Our algorithm takes a collection of textured shapes belonging to the same category and sparse annotations of foreground segments, and produces a joint dense segmentation of the shapes in the collection. We model the segments by a collectively trained Gaussian mixture model. The final model segmentation is formulated as an energy minimization across all models jointly, where intra-model edges control the smoothness and separation of model segments, and inter-model edges impart global consistency. We show promising results on two large real-world datasets, and also compare with previous shape-only 3D segmentation methods using publicly available datasets.

****

How to Evaluate Foreground Maps?

Ran Margolin, Lihi Zelnik-Manor, Ayellet Tal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 248-255

The output of many algorithms in computer-vision is either non-binary maps or binary maps (e.g., salient object detection and object segmentation). Several measures have been suggested to evaluate the accuracy of these foreground maps. In this paper, we show that the most commonly-used measures for evaluating both non-binary maps and binary maps do not always provide a reliable evaluation. This includes the Area-Under-the-Curve measure, the Average-Precision measure, the F-measure, and the evaluation measure of the PASCAL VOC segmentation challenge. We start by identifying three causes of inaccurate evaluation. We then propose a new measure that amends these flaws. An appealing property of our measure is being an intuitive generalization of the F-measure. Finally we propose four meta-measures to compare the adequacy of evaluation measures. We show via experiments that our novel measure is preferable.

****

MILCut: A Sweeping Line Multiple Instance Learning Paradigm for Interactive Image Segmentation

Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, Zhuowen Tu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 256-263

Interactive segmentation, in which a user provides a bounding box to an object of interest for image segmentation, has been applied to a variety of applications in image editing, crowdsourcing, computer vision, and medical imaging. The challenge of this semi-automatic image segmentation task lies in dealing with the uncertainty of the foreground object within a bounding box. Here, we formulate the interactive segmentation problem as a multiple instance learning (MIL) task by generating positive bags from pixels of sweeping lines within a bounding box. We name this approach MILCut. We provide a justification to our formulation and develop an algorithm with significant performance and efficiency gain over existing state-of-the-art systems. Extensive experiments demonstrate the evident advantage of our approach.

****

SCAMS: Simultaneous Clustering and Model Selection

Zhuwen Li, Loong-Fah Cheong, Steven Zhiying Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 264-271

While clustering has been well studied in the past decade, model selection has drawn less attention. This paper addresses both problems in a joint manner with an indicator matrix formulation, in which the clustering cost is penalized by a Frobenius inner product term and the group number estimation is achieved by a rank minimization. As affinity graphs generally contain positive edge values, a sparsity term is further added to avoid the trivial solution. Rather than adopting the conventional convex relaxation approach wholesale, we represent the original problem more faithfully by taking full advantage of the particular structure present in the optimization problem and solving it efficiently using the Alternating Direction Method of Multipliers. The highly constrained nature of the optimization provides our algorithm with the robustness to deal with the varying and often imperfect input affinity matrices arising from different applications and different group numbers. Evaluations on the synthetic data as well as two real wor ld problems show the superiority of the method across a large variety of settings.

****

The Shape-Time Random Field for Semantic Video Labeling

Andrew Kae, Benjamin Marlin, Erik Learned-Miller; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 272-279

We propose a novel discriminative model for semantic labeling in videos by incorporating a prior to model both the shape and temporal dependencies of an object in video. A typical approach for this task is the conditional random field (CRF), which can model local interactions among adjacent regions in a video frame. Recent work has shown how to incorporate a shape prior into a CRF for improving labeling performance, but it may be difficult to model temporal dependencies present in video by using this prior. The conditional restricted Boltzmann machine (CRBM) can model both shape and temporal dependencies, and has been used to learn walking styles from motion-capture data. In this work, we incorporate a CRBM prior into a CRF framework and present a new state-of-the-art model for the task of semantic labeling in videos. In particular, we explore the task of labeling parts of complex face scenes from videos in the YouTube Faces Database (YFDB). Our combined model outperforms competitive baselines both qualitatively and quantitatively.

****

The Secrets of Salient Object Segmentation

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 280-287

In this paper we provide an extensive evaluation of fixation prediction and salient object segmentation algorithms as well as statistics of major datasets. Our analysis identifies serious design flaws of existing salient object benchmarks, called the dataset design bias, by over emphasising the stereotypical concepts of saliency. The dataset design bias does not only create the discomforting disconnection between fixations and salient object segmentation, but also misleads the algorithm designing. Based on our analysis, we propose a new high quality dataset that offers both fixation and salient object segmentation ground-truth. With fixations and salient object being presented simultaneously, we are able to bridge the gap between fixations and salient objects, and propose a novel method for salient object segmentation. Finally, we report significant benchmark progress on 3 existing datasets of segmenting salient objects.

****

Non-rigid Segmentation using Sparse Low Dimensional Manifolds and Deep Belief Networks

Jacinto C. Nascimento, Gustavo Carneiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 288-295

In this paper, we propose a new methodology for segmenting non-rigid visual objects, where the search procedure is onducted directly on a sparse low-dimensional manifold, guided by the classification results computed from a deep belief network. Our main contribution is the fact that we do not rely on the typical sub-division of segmentation tasks into rigid detection and non-rigid delineation. Instead, the non-rigid segmentation is performed directly, where points in the sparse low-dimensional can be mapped to an explicit contour representation in image space. Our proposal shows significantly smaller search and training complexities given that the dimensionality of the manifold is much smaller than the dimensionality of the search spaces for rigid detection and non-rigid delineation aforementioned, and that we no longer require a two-stage segmentation process. We focus on the problem of left ventricle endocardial segmentation from ultrasound images, and lip segmentation from frontal facial images using the extended Cohn-Kanade (CK+) database. Our experiments show that the use of sparse low dimensional manifolds reduces the search and training complexities of current segmentation approaches without a significant impact on the segmentation accuracy shown by state-of-the-art approaches.

****

An Exemplar-based CRF for Multi-instance Object Segmentation
Xuming He, Stephen Gould; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 296-303

We address the problem of joint detection and segmentation of multiple object instances in an image, a key step towards scene understanding. Inspired by data-driven methods, we propose an exemplar-based approach to the task of instance segmentation, in which a set of reference image/shape masks is used to find multiple objects. We design a novel CRF framework that jointly models object appearance, shape deformation, and object occlusion. To tackle the challenging MAP inference problem, we derive an alternating procedure that interleaves object segmentation and shape/appearance adaptation. We evaluate our method on two datasets with instance labels and show promising results.

****

Object Partitioning using Local Convexity
Simon Christoph Stein, Markus Schoeler, Jeremie Papon, Florentin Worgotter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 304-311

The problem of how to arrive at an appropriate 3D-segmentation of a scene remains difficult. While current state-of-the-art methods continue to gradually improve in benchmark performance, they also grow more and more complex, for example by incorporating chains of classifiers, which require training on large manually annotated data-sets. As an alternative to this, we present a new, efficient learning- and model-free approach for the segmentation of 3D point clouds into object parts. The algorithm begins by decomposing the scene into an adjacency-graph of surface patches based on a voxel grid. Edges in the graph are then classified as either convex or concave using a novel combination of simple criteria which operate on the local geometry of these patches. This way the graph is divided into locally convex connected subgraphs, which -- with high accuracy -- represent object parts. Additionally, we propose a novel depth dependent voxel grid to deal with the decreasing point-density at far distances in the point clouds. This improves segmentation, allowing the use of fixed parameters for vastly different scenes. The algorithm is straightforward to implement and requires no training data, while nevertheless producing results that are comparable to state-of-the-art methods which incorporate high-level concepts involving classification, learning and model fitting.

****

Bayesian Active Contours with Affine-Invariant, Elastic Shape Prior
Darshan Bryner, Anuj Srivastava; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 312-319

Active contour, especially in conjunction with prior-shape models, has become an important tool in image segmentation. However, most contour methods use shape priors based on similarity-shape analysis, i.e. analysis that is invariant to rotation, translation, and scale. In practice, the training shapes used for prior-shape models may be collected from viewing angles different from those for the test images and require invariance to a larger class of transformation. Using an elastic, affine-invariant shape modeling of planar curves, we propose an active contour algorithm in which the training and test shapes can be at arbitrary affine transformations, and the resulting segmentation is robust to perspective skews. We construct a shape space of affine-standardized curves and derive a statistical model for capturing class-specific shape variability. The active contour is then driven by the true gradient of a total energy composed of a data term, a smoothing term, and an affine-invariant shape-prior term. This framework is demonstrated using a number of examples involving the segmentation of occluded or noisy images of targets subject to perspective skew.

****

Max-Margin Boltzmann Machines for Object Segmentation
Jimei Yang, Simon Safar, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 320-327

We present Max-Margin Boltzmann Machines (MMBMs) for object segmentation. MMBMs are essentially a class of Conditional Boltzmann Machines that model the joint d istribution of hidden variables and output labels conditioned on input observations. In addition to image-to-label connections, we build direct image-to-hidden connections to facilitate global shape prediction, and thus derive a simple Iterated Conditional Modes algorithm for efficient maximum a posteriori inference. We formulate a max-margin objective function for discriminative training, and analyze the effects of different margin functions on learning. We evaluate MMBMs using three datasets against state-of-the-art methods to demonstrate the strength of the proposed algorithms.

****

Multiscale Combinatorial Grouping
Pablo Arbelaez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marques, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 328-335

We propose a unified approach for bottom-up hierarchical image segmentation and object candidate generation for recognition, called Multiscale Combinatorial Grouping (MCG). For this purpose, we first develop a fast normalized cuts algorithm. We then propose a high-performance hierarchical segmenter that makes effective use of multiscale information. Finally, we propose a grouping strategy that combines our multiscale regions into highly-accurate object candidates by exploring efficiently their combinatorial space. We conduct extensive experiments on both the BSDS500 and on the PASCAL 2012 segmentation datasets, showing that MCG produces state-of-the-art contours, hierarchical regions and object candidates.

****

RIGOR: Reusing Inference in Graph Cuts for Generating Object Regions
Ahmad Humayun, Fuxin Li, James M. Rehg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 336-343

Popular figure-ground segmentation algorithms generate a pool of boundary-aligned segment proposals that can be used in subsequent object recognition engines. These algorithms can recover most image objects with high accuracy, but are usually computationally intensive since many graph cuts are computed with different enumerations of segment seeds. In this paper we propose an algorithm, RIGOR, for efficiently generating a pool of overlapping segment proposals in images. By precomputing a graph which can be used for parametric min-cuts over different seeds, we speed up the generation of the segment pool. In addition, we have made design choices that avoid extensive computations without losing performance. In particular, we demonstrate that the segmentation performance of our algorithm is slightly better than the state-of-the-art on the PASCAL VOC dataset, while being an order of magnitude faster.

****

Efficient Hierarchical Graph-Based Segmentation of RGBD Videos
Steven Hickson, Stan Birchfield, Irfan Essa, Henrik Christensen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 344-351

We present an efficient and scalable algorithm for segmenting 3D RGBD point clouds by combining depth, color, and temporal information using a multistage, hierarchical graph-based approach. Our algorithm processes a moving window over several point clouds to group similar regions over a graph, resulting in an initial over-segmentation. These regions are then merged to yield a dendrogram using agglomerative clustering via a minimum spanning tree algorithm. Bipartite graph matching at a given level of the hierarchical tree yields the final segmentation of the point clouds by maintaining region identities over arbitrarily long periods of time. We show that a multistage segmentation with depth then color yields better results than a linear combination of depth and color. Due to its incremental processing, our algorithm can process videos of any length and in a streaming pipeline. The algorithm's ability to produce robust, efficient segmentation is demonstrated with numerous experimental results on challenging sequences from our own as well as public RGBD data sets.

****

Point Matching in the Presence of Outliers in Both Point Sets: A Concave Optimization Approach

Wei Lian, Lei Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 352-359

Recently, a concave optimization approach has been proposed to solve the robust point matching (RPM) problem. This method is globally optimal, but it requires that each model point has a counterpart in the data point set. Unfortunately, such a requirement may not be satisfied in certain applications when there are outliers in both point sets. To address this problem, we relax this condition and reduce the objective function of RPM to a function with few nonlinear terms by eliminating the transformation variables. The resulting function, however, is no longer quadratic. We prove that it is still concave over the feasible region of point correspondence. The branch-and-bound (BnB) algorithm can then be used for optimization. To further improve the efficiency of the BnB algorithm whose bottleneck lies in the costly computation of the lower bound, we propose a new lower bounding scheme which has a k-cardinality linear assignment formulation and can be efficiently solved. Experimental results show that the proposed algorithm outperforms state-of-the-arts in its robustness to disturbances and point matching accuracy.

****

Multiple Structured-Instance Learning for Semantic Segmentation with Uncertain Training Data

Feng-Ju Chang, Yen-Yu Lin, Kuang-Jui Hsu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 360-367

We present an approach MSIL-CRF that incorporates multiple instance learning (MIL) into conditional random fields (CRFs). It can generalize CRFs to work on training data with uncertain labels by the principle of MIL. In this work, it is applied to saving manual efforts on annotating training data for semantic segmentation. Specifically, we consider the setting in which the training dataset for semantic segmentation is a mixture of a few object segments and an abundant set of objects' bounding boxes. Our goal is to infer the unknown object segments enclosed by the bounding boxes so that they can serve as training data for semantic segmentation. To this end, we generate multiple segment hypotheses for each bounding box with the assumption that at least one hypothesis is close to the ground truth. By treating a bounding box as a bag with its segment hypotheses as structured instances, MSIL-CRF selects the most likely segment hypotheses by leveraging the knowledge derived from both the labeled and uncertain training data. The experimental results on the Pascal VOC segmentation task demonstrate that MSIL-CRF can provide effective alternatives to manually labeled segments for semantic segmentation.

****

Joint Motion Segmentation and Background Estimation in Dynamic Scenes

Adeel Mumtaz, Weichen Zhang, Antoni B. Chan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 368-375

We propose a joint foreground-background mixture model (FBM) that simultaneously performs background estimation and motion segmentation in complex dynamic scenes. Our FBM consist of a set of location-specific dynamic texture (DT) components, for modeling local background motion, and set of global DT components, for modeling consistent foreground motion. We derive an EM algorithm for estimating the parameters of the FBM. We also apply spatial constraints to the FBM using an Markov random field grid, and derive a corresponding variational approximation for inference. Unlike existing approaches to background subtraction, our FBM does not require a manually selected threshold or a separate training video. Unlike existing motion segmentation techniques, our FBM can segment foreground motions over complex background with mixed motions, and detect stopped objects. Since most dynamic scene datasets only contain videos with a single foreground object over a simple background, we develop a new challenging dataset with multiple foreground objects over complex dynamic backgrounds. In experiments, we show that jointly modeling the background and foreground segments with FBM yields significant improvements in accuracy on both background estimation and motion segmentation, compared to state-of-the-art methods.

****

SeamSeg: Video Object Segmentation using Patch Seams
S. Avinash Ramakanth, R. Venkatesh Babu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 376-383
In this paper, we propose a technique for video object segmentation using patch seams across frames. Typically, seams, which are connected paths of low energy, are utilised for retargeting, where the primary aim is to reduce the image size while preserving the salient image contents. Here, we adapt the formulation of seams for temporal label propagation. The energy function associated with the proposed video seams provides temporal linking of patches across frames, to accurately segment the object. The proposed energy function takes into account the similarity of patches along the seam, temporal consistency of motion and spatial coherency of seams. Label propagation is achieved with high fidelity in the critical boundary regions, utilising the proposed patch seams. To achieve this without additional overheads, we curtail the error propagation by formulating boundary regions as rough-sets. The proposed approach out-perform state-of-the-art supervised and unsupervised algorithms, on benchmark datasets.
****
Laplacian Coordinates for Seeded Image Segmentation
Wallace Casaca, Luis Gustavo Nonato, Gabriel Taubin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 384-391
Seed-based image segmentation methods have gained much attention lately, mainly due to their good performance in segmenting complex images with little user interaction. Such popularity leveraged the development of many new variations of seed-based image segmentation techniques, which vary greatly regarding mathematical formulation and complexity. Most existing methods in fact rely on complex mathematical formulations that typically do not guarantee unique solution for the segmentation problem while still being prone to be trapped in local minima. In this work we present a novel framework for seed-based image segmentation that is mathematically simple, easy to implement, and guaranteed to produce a unique solution. Moreover, the formulation holds an anisotropic behavior, that is, pixels sharing similar attributes are kept closer to each other while big jumps are naturally imposed on the boundary between image regions, thus ensuring better fitting on object boundaries. We show that the proposed framework outperform state-of-the-art techniques in terms of quantitative quality metrics as well as qualitative visual results.
****
Error-tolerant Scribbles Based Interactive Image Segmentation
Junjie Bai, Xiaodong Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 392-399
Scribbles in scribble-based interactive segmentation such as graph-cut are usually assumed to be perfectly accurate, i.e., foreground scribble pixels will never be segmented as background in the final segmentation. However, it can be hard to draw perfectly accurate scribbles, especially on fine structures of the image or on mobile touch-screen devices. In this paper, we propose a novel ratio energy function that tolerates errors in the user input while encouraging maximum use of the user input information. More specifically, the ratio energy aims to minimize the graph-cut energy while maximizing the user input respected in the segmentation. The ratio energy function can be exactly optimized using an efficient iterated graph cut algorithm. The robustness of the proposed method is validated on the GrabCut dataset using both synthetic scribbles and manual scribbles. The experimental results show that the proposed algorithm is robust to the errors in the user input and preserves the "anchoring" capability of the user input.
****
Iterative Multilevel MRF Leveraging Context and Voxel Information for Brain Tumour Segmentation in MRI
Nagesh Subbanna, Doina Precup, Tal Arbel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 400-405
In this paper, we introduce a fully automated multistage graphical probabilistic framework to segment brain tumours from multimodal Magnetic Resonance Images (MRIs) acquired from real patients. An initial Bayesian tumour classification base

d on Gabor texture features permits subsequent computations to be focused on areas where the probability of tumour is deemed high. An iterative, multistage Markov Random Field (MRF) framework is then devised to classify the various tumour subclasses (e.g. edema, solid tumour, enhancing tumour and necrotic core). Specifically, an adapted, voxel-based MRF provides tumour candidates to a higher level, regional MRF, which then leverages both contextual texture information and relative spatial consistency of the tumour subclass positions to provide updated regional information down to the voxel-based MRF for further local refinement. The two stages iterate until convergence. Experiments are performed on publicly available, patient brain tumour images from the MICCAI 2012 [11] and 2013 [12] Brain Tumour Segmentation Challenges. The results demonstrate that the proposed method achieves the top performance in the segmentation of tumour cores and enhancing tumours, and performs comparably to the winners in other tumour categories.

****

Large Scale Multi-view Stereopsis Evaluation

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, Henrik Aanaes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 406-413

The seminal multiple view stereo benchmark evaluations from Middlebury and by Strecha et al. have played a major role in propelling the development of multi-view stereopsis methodology. Although seminal, these benchmark datasets are limited in scope with few reference scenes. Here, we try to take these works a step further by proposing a new multi-view stereo dataset, which is an order of magnitude larger in number of scenes and with a significant increase in diversity. Specifically, we propose a dataset containing 80 scenes of large variability. Each scene consists of 49 or 64 accurate camera positions and reference structured light scans, all acquired by a 6-axis industrial robot. To apply this dataset we propose an extension of the evaluation protocol from the Middlebury evaluation, reflecting the more complex geometry of some of our scenes. The proposed dataset is used to evaluate the state of the art multiview stereo algorithms of Tola et al., Campbell et al. and Furukawa et al. Hereby we demonstrate the usability of the dataset as well as gain insight into the workings and challenges of multi-view stereopsis. Through these experiments we empirically validate some of the central hypotheses of multi-view stereopsis, as well as determining and reaffirming some of the central challenges.

****

Timing-Based Local Descriptor for Dynamic Surfaces

Tony Tung, Takashi Matsuyama; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 414-421

In this paper, we present the first local descriptor designed for dynamic surfaces. A dynamic surface is a surface that can undergo non-rigid deformation (e.g., human body surface). Using state-of-the-art technology, details on dynamic surfaces such as cloth wrinkle or facial expression can be accurately reconstructed. Hence, various results (e.g., surface rigidity, or elasticity) could be derived by microscopic categorization of surface elements. We propose a timing-based descriptor to model local spatiotemporal variations of surface intrinsic properties. The low-level descriptor encodes gaps between local event dynamics of neighboring keypoints using timing structure of linear dynamical systems (LDS). We also introduce the bag-of-timings (BoT) paradigm for surface dynamics characterization. Experiments are performed on synthesized and real-world datasets. We show the proposed descriptor can be used for challenging dynamic surface classification and segmentation with respect to rigidity at surface keypoints.

****

A Minimal Solution to the Generalized Pose-and-Scale Problem

Jonathan Ventura, Clemens Arth, Gerhard Reitmayr, Dieter Schmalstieg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 422-429

We propose a novel solution to the generalized camera pose problem which includes the internal scale of the generalized camera as an unknown parameter. This further generalization of the well-known absolute camera pose problem has applicat ions in multi-frame loop closure. While a well-calibrated camera rig has a fixed and known scale, camera trajectories produced by monocular motion estimation necessarily lack a scale estimate.   Thus, when performing loop closure in monocular visual odometry, or registering separate structure-from-motion reconstructions, we must estimate a seven degree-of-freedom similarity transform from corresponding observations. Existing approaches solve this problem, in specialized configurations, by aligning 3D triangulated points or individual camera pose estimates. Our approach handles general configurations of rays and points and directly estimates the full similarity transformation from the 2D-3D correspondences. Four correspondences are needed in the minimal case, which has eight possible solutions. The minimal solver can be used in a hypothesize-and-test architecture for robust transformation estimation. Our solver also produces a least-squares estimate in the overdetermined case. The approach is evaluated experimentally on synthetic and real datasets, and is shown to produce higher accuracy solutions to multi-frame loop closure than existing approaches.

****

A General and Simple Method for Camera Pose and Focal Length Determination

Yinqiang Zheng, Shigeki Sugimoto, Imari Sato, Masatoshi Okutomi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 430-437

In this paper, we revisit the pose determination problem of a partially calibrated camera with unknown focal length, hereafter referred to as the PnPf problem, by using n (n ≥ 4) 3D-to-2D point correspondences. Our core contribution is to introduce the angle constraint and derive a compact bivariate polynomial equation for each point triplet. Based on this polynomial equation, we propose a truly general method for the PnPf problem, which is suited both to the minimal 4-point based RANSAC application, and also to large scale scenarios with thousands of points, irrespective of the 3D point configuration. In addition, by solving bivariate polynomial systems via the Sylvester resultant, our method is very simple and easy to implement. Its simplicity is especially obvious when one needs to develop a fast solver for the 4-point case on the basis of the characteristic polynomial technique. Experiment results have also demonstrated its superiority in accuracy and efficiency when compared with the existing state-of-the-art solutions.

****

Partial Symmetry in Polynomial Systems and its Applications in Computer Vision

Yubin Kuang, Yinqiang Zheng, Kalle Astrom; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 438-445

Algorithms for solving systems of polynomial equations are key components for solving geometry problems in computer vision. Fast and stable polynomial solvers are essential for numerous applications e.g. minimal problems or finding for all stationary points of certain algebraic errors. Recently, full symmetry in the polynomial systems has been utilized to simplify and speed up state-of-the-art polynomial solvers based on Grobner basis method. In this paper, we further explore partial symmetry (i.e. where the symmetry lies in a subset of the variables) in the polynomial systems. We develop novel numerical schemes to utilize such partial symmetry. We then demonstrate the advantage of our schemes in several computer vision problems. In both synthetic and real experiments, we show that utilizing partial symmetry allow us to obtain faster and more accurate polynomial solvers than the general solvers.

****

Efficient Computation of Relative Pose for Multi-Camera Systems

Laurent Kneip, Hongdong Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 446-453

We present a novel solution to compute the relative pose of a generalized camera. Existing solutions are either not general, have too high computational complexity, or require too many correspondences, which impedes an efficient or accurate usage within Ransac schemes. We factorize the problem as a low-dimensional, iterative optimization over relative rotation only, directly derived from well-known epipolar constraints. Common generalized cameras often consist of camera clusters, and give rise to omni-directional landmark observations. We prove that our

iterative scheme performs well in such practically relevant situations, eventually resulting in computational efficiency similar to linear solvers, and accuracy close to bundle adjustment, while using less correspondences. Experiments on both virtual and real multi-camera systems prove superior overall performance for robust, real-time multi-camera motion-estimation.
****

Simultaneous Localization and Calibration: Self-Calibration of Consumer Depth Cameras

Qian-Yi Zhou, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 454-460

We describe an approach for simultaneous localization and calibration of a stream of range images. Our approach jointly optimizes the camera trajectory and a calibration function that corrects the camera's unknown nonlinear distortion. Experiments with real-world benchmark data and synthetic data show that our approach increases the accuracy of camera trajectories and geometric models estimated from range video produced by consumer-grade cameras.
****

Minimal Scene Descriptions from Structure from Motion Models

Song Cao, Noah Snavely; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 461-468

How much data do we need to describe a location? We explore this question in the context of 3D scene reconstructions created from running structure from motion on large Internet photo collections, where reconstructions can contain many millions of 3D points. We consider several methods for computing much more compact representations of such reconstructions for the task of location recognition, with the goal of maintaining good performance with very small models. In particular, we introduce a new method for computing compact models that takes into account both image-point relationships and feature distinctiveness, and we show that this method produces small models that yield better recognition performance than previous model reduction techniques.
****

Fast, Approximate Piecewise-Planar Modeling Based on Sparse Structure-from-Motion and Superpixels

Andras Bodis-Szomoru, Hayko Riemenschneider, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 469-476

State-of-the-art Multi-View Stereo (MVS) algorithms deliver dense depth maps or complex meshes with very high detail, and redundancy over regular surfaces. In turn, our interest lies in an approximate, but light-weight method that is better to consider for large-scale applications, such as urban scene reconstruction from ground-based images. We present a novel approach for producing dense reconstructions from multiple images and from the underlying sparse Structure-from-Motion (SfM) data in an efficient way. To overcome the problem of SfM sparsity and textureless areas, we assume piecewise planarity of man-made scenes and exploit both sparse visibility and a fast over-segmentation of the images. Reconstruction is formulated as an energy-driven, multi-view plane assignment problem, which we solve jointly over superpixels from all views while avoiding expensive photoconsistency computations. The resulting planar primitives -- defined by detailed superpixel boundaries -- are computed in about 10 seconds per image.
****

On Projective Reconstruction In Arbitrary Dimensions

Behrooz Nasihatkon, Richard Hartley, Jochen Trumpf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 477-484

We study the theory of projective reconstruction for multiple projections from an arbitrary dimensional projective space into lower-dimensional spaces. This problem is important due to its applications in the analysis of dynamical scenes. The current theory, due to Hartley and Schaffalitzky, is based on the Grassmann tensor, generalizing the ideas of fundamental matrix, trifocal tensor and quadrifocal tensor used in the well-studied case of 3D to 2D projections. We present a theory whose point of departure is the projective equations rather than the Gras smann tensor. This is a better fit for the analysis of approaches such as bundle adjustment and projective factorization which seek to directly solve the projective equations. In a first step, we prove that there is a unique Grassmann tensor corresponding to each set of image points, a question that remained open in the work of Hartley and Schaffalitzky. Then, we prove that projective equivalence follows from the set of projective equations given certain conditions on the estimated camera-point setup or the estimated projective depths. Finally, we demonstrate how wrong solutions to the projective factorization problem can happen, and classify such degenerate solutions based on the zero patterns in the estimated depth matrix.

****

Stereo under Sequential Optimal Sampling: A Statistical Analysis Framework for Search Space Reduction

Yilin Wang, Ke Wang, Enrique Dunn, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 485-492

We develop a sequential optimal sampling framework for stereo disparity estimation by adapting the Sequential Probability Ratio Test (SPRT) model. We operate over local image neighborhoods by iteratively estimating single pixel disparity values until sufficient evidence has been gathered to either validate or contradict the current hypothesis regarding local scene structure. The output of our sampling is a set of sampled pixel positions along with a robust and compact estimate of the set of disparities contained within a given region. We further propose an efficient plane propagation mechanism that leverages the pre-computed sampling positions and the local structure model described by the reduced local disparity set. Our sampling framework is a general pre-processing mechanism aimed at reducing computational complexity of disparity search algorithms by ascertaining a reduced set of disparity hypotheses for each pixel. Experiments demonstrate the effectiveness of the proposed approach when compared to state of the art methods.

****

Efficient Pruning LMI Conditions for Branch-and-Prune Rank and Chirality-Constrained Estimation of the Dual Absolute Quadric

Adlane Habed, Danda Pani Paudel, Cedric Demonceaux, David Fofi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 493-500

We present a new globally optimal algorithm for self-calibrating a moving camera with constant parameters. Our method aims at estimating the Dual Absolute Quadric (DAQ) under the rank-3 and, optionally, camera centers chirality constraints. We employ the Branch-and-Prune paradigm and explore the space of only 5 parameters. Pruning in our method relies on solving Linear Matrix Inequality (LMI) feasibility and Generalized Eigenvalue (GEV) problems that solely depend upon the entries of the DAQ. These LMI and GEV problems are used to rule out branches in the search tree in which a quadric not satisfying the rank and chirality conditions on camera centers is guaranteed not to exist. The chirality LMI conditions are obtained by relying on the mild assumption that the camera undergoes a rotation of no more than 90 degrees between consecutive views. Furthermore, our method does not rely on calculating bounds on any particular cost function and hence can virtually optimize any objective while achieving global optimality in a very competitive running-time.

****

Very Fast Solution to the PnP Problem with Algebraic Outlier Rejection

Luis Ferraz, Xavier Binefa, Francesc Moreno-Noguer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 501-508

We propose a real-time, robust to outliers and accurate solution to the Perspective-n-Point (PnP) problem. The main advantages of our solution are twofold: first, it in- tegrates the outlier rejection within the pose estimation pipeline with a negligible computational overhead; and sec- ond, its scalability to arbitrarily large number of correspon- dences. Given a set of 3D-to-2D matches, we formulate pose estimation problem as a low-rank homogeneous sys- tem where the solution lies on its 1D null space. Outlier correspondences are those rows of the line

ar system which perturb the null space and are progressively detected by projecting them on an iteratively estimated solution of the null space. Since our outlier removal process is based on an algebraic criterion which does not require computing the full-pose and reprojecting back all 3D points on the image plane at each step, we achieve speed gains of more than 100Ã■ compared to RANSAC strategies. An extensive exper- imental evaluation will show that our solution yields accu- rate results in situations with up to 50% of outliers, and can process more than 1000 correspondences in less than 5ms.
****

Finding Vanishing Points via Point Alignments in Image Primal and Dual Domains
Jose Lezama, Rafael Grompone von Gioi, Gregory Randall, Jean-Michel Morel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 509-515

We present a novel method for automatic vanishing point detection based on primal and dual point alignment detection. The very same point alignment detection algorithm is used twice: First in the image domain to group line segment endpoints into more precise lines. Second, it is used in the dual domain where converging lines become aligned points. The use of the recently introduced PClines dual spaces and a robust point alignment detector leads to a very accurate algorithm. Experimental results on two public standard datasets show that our method significantly advances the state-of-the-art in the Manhattan world scenario, while producing state-of-the-art performances in non-Manhattan scenes.
****

Discriminative Feature-to-Point Matching in Image-Based Localization
Michael Donoser, Dieter Schmalstieg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 516-523

The prevalent approach to image-based localization is matching interest points detected in the query image to a sparse 3D point cloud representing the known world. The obtained correspondences are then used to recover a precise camera pose. The state-of-the-art in this field often ignores the availability of a set of 2D descriptors per 3D point, for example by representing each 3D point by only its centroid. In this paper we demonstrate that these sets contain useful information that can be exploited by formulating matching as a discriminative classification problem. Since memory demands and computational complexity are crucial in such a setup, we base our algorithm on the efficient and effective random fern principle. We propose an extension which projects features to fern-specific embedding spaces, which yields improved matching rates in short runtime. Experiments first show that our novel formulation provides improved matching performance in comparison to the standard nearest neighbor approach and that we outperform related randomization methods in our localization scenario.
****

Two-View Camera Housing Parameters Calibration for Multi-Layer Flat Refractive Interface
Xida Chen, Yee-Hong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 524-531

In this paper, we present a novel refractive calibration method for an underwater stereo camera system where both cameras are looking through multiple parallel flat refractive interfaces. At the heart of our method is an important finding that the thickness of the interface can be estimated from a set of pixel correspondences in the stereo images when the refractive axis is given. To our best knowledge, such a finding has not been studied or reported. Moreover, by exploring the search space for the refractive axis and using reprojection error as a measure, both the refractive axis and the thickness of the interface can be recovered simultaneously. Our method does not require any calibration target such as a checkerboard pattern which may be difficult to manipulate when the cameras are deployed deep undersea. The implementation of our method is simple. In particular, it only requires solving a set of linear equations of the form Ax = b and applies sparse bundle adjustment to refine the initial estimated results. Extensive experiments have been carried out which include simulations with and without outliers to verify the correctness of our method as well as to test its robustness to noise and outliers. The results of real experiments are also provided. The accuracy of our results is comparable to that of a state-of-the-art method that requires known 3D geometry of a scene.

****

Accurate Localization and Pose Estimation for Large 3D Models

Linus Svarm, Olof Enqvist, Magnus Oskarsson, Fredrik Kahl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 532-539

We consider the problem of localizing a novel image in a large 3D model. In principle, this is just an instance of camera pose estimation, but the scale introduces some challenging problems. For one, it makes the correspondence problem very difficult and it is likely that there will be a significant rate of outliers to handle.   In this paper we use recent theoretical as well as technical advances to tackle these problems. Many modern cameras and phones have gravitational sensors that allow us to reduce the search space. Further, there are new techniques to efficiently and reliably deal with extreme rates of outliers. We extend these methods to camera pose estimation by using accurate approximations and fast polynomial solvers. Experimental results are given demonstrating that it is possible to reliably estimate the camera pose despite more than 99% of outlier correspondences.

****

Relative Pose Estimation for a Multi-Camera System with Known Vertical Direction

Gim Hee Lee, Marc Pollefeys, Friedrich Fraundorfer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 540-547

In this paper, we present our minimal 4-point and linear 8-point algorithms to estimate the relative pose of a multi-camera system with known vertical directions, i.e. known absolute roll and pitch angles. We solve the minimal 4-point algorithm with the hidden variable resultant method and show that it leads to an 8-degree univariate polynomial that gives up to 8 real solutions. We identify a degenerated case from the linear 8-point algorithm when it is solved with the standard Singular Value Decomposition (SVD) method and adopt a simple alternative solution which is easy to implement. We show that our proposed algorithms can be efficiently used within RANSAC for robust estimation. We evaluate the accuracy of our proposed algorithms by comparisons with various existing algorithms for the multi-camera system on simulations and show the feasibility of our proposed algorithms with results from multiple real-world datasets.

****

Optimal Decisions from Probabilistic Models: The Intersection-over-Union Case

Sebastian Nowozin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 548-555

A probabilistic model allows us to reason about the world and make statistically optimal decisions using Bayesian decision theory. However, in practice the intractability of the decision problem forces us to adopt simplistic loss functions such as the 0/1 loss or Hamming loss and as result we make poor decisions through MAP estimates or through low-order marginal statistics. In this work we investigate optimal decision making for more realistic loss functions. Specifically we consider the popular intersection-over-union (IoU) score used in image segmentation benchmarks and show that it results in a hard combinatorial decision problem. To make this problem tractable we propose a statistical approximation to the objective function, as well as an approximate algorithm based on parametric linear programming. We apply the algorithm on three benchmark datasets and obtain improved intersection-over-union scores compared to maximum-posterior-marginal decisions. Our work points out the difficulties of using realistic loss functions with probabilistic computer vision models.

****

Covariance Trees for 2D and 3D Processing

Thierry Guillemot, Andres Almansa, Tamy Boubekeur; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 556-563

Gaussian Mixture Models have become one of the major tools in modern statistical image processing, and allowed performance breakthroughs in patch-based image de noising and restoration problems. Nevertheless, their adoption level was kept relatively low because of the computational cost associated to learning such models on large image databases. This work provides a flexible and generic tool for dealing with such models without the computational penalty or parameter tuning difficulties associated to a naÃ¯ve implementation of GMM-based image restoration tasks. It does so by organising the data manifold in a hirerachical multiscale structure (the Covariance Tree) that can be queried at various scale levels around any point in feature-space. We start by explaining how to construct a Covariance Tree from a subset of the input data, how to enrich its statistics from a larger set in a streaming process, and how to query it efficiently, at any scale. We then demonstrate its usefulness on several applications, including non-local image filtering, data-driven denoising, reconstruction from random samples and surface modeling from unorganized 3D points sets.

****

Hierarchical Subquery Evaluation for Active Learning on a Graph

Oisin Mac Aodha, Neill D.F. Campbell, Jan Kautz, Gabriel J. Brostow; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 564-571

To train good supervised and semi-supervised object classifiers, it is critical that we not waste the time of the human experts who are providing the training labels. Existing active learning strategies can have uneven performance, being efficient on some datasets but wasteful on others, or inconsistent just between runs on the same dataset. We propose perplexity based graph construction and a new hierarchical subquery evaluation algorithm to combat this variability, and to release the potential of Expected Error Reduction. Under some specific circumstances, Expected Error Reduction has been one of the strongest-performing informativeness criteria for active learning. Until now, it has also been prohibitively costly to compute for sizeable datasets. We demonstrate our highly practical algorithm, comparing it to other active learning measures on classification datasets that vary in sparsity, dimensionality, and size. Our algorithm is consistent over multiple runs and achieves high accuracy, while querying the human expert for labels at a frequency that matches their desired time budget.

****

Anytime Recognition of Objects and Scenes

Sergey Karayev, Mario Fritz, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 572-579

Humans are capable of perceiving a scene at a glance, and obtain deeper understanding with additional time. Similarly, visual recognition deployments should be robust to varying computational budgets. Such situations require Anytime recognition ability, which is rarely considered in computer vision research. We present a method for learning dynamic policies to optimize Anytime performance in visual architectures. Our model sequentially orders feature computation and performs subsequent classification. Crucially, decisions are made at test time and depend on observed data and intermediate results. We show the applicability of this system to standard problems in scene and object recognition. On suitable datasets, we can incorporate a semantic back-off strategy that gives maximally specific predictions for a desired level of accuracy; this provides a new view on the time course of human visual perception.

****

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580-587

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012---achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural netw orks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since we combine region proposals with CNNs, we call our method R-CNN: Regions with CNN features. We also present experiments that provide insight into what the network learns, revealing a rich hierarchy of image features. Source code for the complete system is available at http://www.cs.berkeley.edu/~rbg/rcnn.

****

Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group

Raviteja Vemulapalli, Felipe Arrate, Rama Chellappa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 588-595

Recently introduced cost-effective depth sensors coupled with the real-time skeleton estimation algorithm of Shotton et al. [16] have generated a renewed interest in skeleton-based human action recognition. Most of the existing skeleton-based approaches use either the joint locations or the joint angles to represent a human skeleton. In this paper, we propose a new skeletal representation that explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space. Since 3D rigid body motions are members of the special Euclidean group SE(3), the proposed skeletal representation lies in the Lie group SE(3)Ã■. . .Ã■SE(3), which is a curved manifold. Using the proposed representation, human actions can be modeled as curves in this Lie group. Since classification of curves in this Lie group is not an easy task, we map the action curves from the Lie group to its Lie algebra, which is a vector space. We then perform classification using a combination of dynamic time warping, Fourier temporal pyramid representation and linear SVM. Experimental results on three action datasets show that the proposed representation performs better than many existing skeletal representations. The proposed approach also outperforms various state-of-the-art skeleton-based human action recognition approaches.

****

Multi-View Super Vector for Action Recognition

Zhuowei Cai, Limin Wang, Xiaojiang Peng, Yu Qiao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 596-603

Images and videos are often characterized by multiple types of local descriptors such as SIFT, HOG and HOF, each of which describes certain aspects of object feature. Recognition systems benefit from fusing multiple types of these descriptors. Two widely applied fusion pipelines are descriptor concatenation and kernel average. The first one is effective when different descriptors are strongly correlated, while the second one is probably better when descriptors are relatively independent. In practice, however, different descriptors are neither fully independent nor fully correlated, and previous fusion methods may not be satisfying. In this paper, we propose a new global representation, Multi-View Super Vector (MVSV), which is composed of relatively independent components derived from a pair of descriptors. Kernel average is then applied on these components to produce recognition result. To obtain MVSV, we develop a generative mixture model of probabilistic canonical correlation analyzers (M-PCCA), and utilize the hidden factors and gradient vectors of M-PCCA to construct MVSV for video representation. Experiments on video based action recognition tasks show that MVSV achieves promising results, and outperforms FV and VLAD with descriptor concatenation or kernel average fusion strategy.

****

Unsupervised Spectral Dual Assignment Clustering of Human Actions in Context

Simon Jones, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 604-611

A recent trend of research has shown how contextual information related to an action, such as a scene or object, can enhance the accuracy of human action recognition systems. However, using context to improve unsupervised human action clustering has never been considered before, and cannot be achieved using existing clustering methods. To solve this problem, we introduce a novel, general purpose algorithm, Dual Assignment k-Means (DAKM), which is uniquely capable of performin g two co-occurring clustering tasks simultaneously, while exploiting the correlation information to enhance both clusterings. Furthermore, we describe a spectral extension of DAKM (SDAKM) for better performance on realistic data. Extensive experiments on synthetic data and on three realistic human action datasets with scene context show that DAKM/SDAKM can significantly outperform the state-of-the-art clustering methods by taking into account the contextual relationship between actions and scenes.

****

Parsing Videos of Actions with Segmental Grammars

Hamed Pirsiavash, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 612-619

Real-world videos of human activities exhibit temporal structure at various scales; long videos are typically composed out of multiple action instances, where each instance is itself composed of sub-actions with variable durations and orderings. Temporal grammars can presumably model such hierarchical structure, but are computationally difficult to apply for long video streams. We describe simple grammars that capture hierarchical temporal structure while admitting inference with a finite-state-machine. This makes parsing linear time, constant storage, and naturally online. We train grammar parameters using a latent structural SVM, where latent subactions are learned automatically. We illustrate the effectiveness of our approach over common baselines on a new half-million frame dataset of continuous YouTube videos.

****

Rate-Invariant Analysis of Trajectories on Riemannian Manifolds with Application in Visual Speech Recognition

Jingyong Su, Anuj Srivastava, Fillipe D. M. de Souza, Sudeep Sarkar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 620-627

In statistical analysis of video sequences for speech recognition, and more generally activity recognition, it is natural to treat temporal evolutions of features as trajectories on Riemannian manifolds. However, different evolution patterns result in arbitrary parameterizations of these trajectories. We investigate a recent framework from statistics literature that handles this nuisance variability using a cost function/distance for temporal registration and statistical summarization & modeling of trajectories. It is based on a mathematical representation of trajectories, termed transported square-root vector field (TSRVF), and the L2 norm on the space of TSRVFs. We apply this framework to the problem of speech recognition using both audio and visual components. In each case, we extract features, form trajectories on corresponding manifolds, and compute parametrization-invariant distances using TSRVFs for speech classification. On the OuluVS database the classification performance under metric increases significantly, by nearly 100% under both modalities and for all choices of features. We obtained speaker-dependent classification rate of 70% and 96% for visual and audio components, respectively.

****

Piecewise Planar and Compact Floorplan Reconstruction from Images

Ricardo Cabral, Yasutaka Furukawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 628-635

This paper presents a system to reconstruct piecewise planar and compact floorplans from images, which are then converted to high quality texture-mapped models for free- viewpoint visualization. There are two main challenges in image-based floorplan reconstruction. The first is the lack of 3D information that can be extracted from images by Structure from Motion and Multi-View Stereo, as indoor scenes abound with non-diffuse and homogeneous surfaces plus clutter. The second challenge is the need of a sophisti- cated regularization technique that enforces piecewise pla- narity, to suppress clutter and yield high quality texture mapped models. Our technical contributions are twofold. First, we propose a novel structure classification technique to classify each pixel to three regions (floor, ceiling, and wall), which provide 3D cues even from a single image. Second, we cast floorplan reconstruction as a shortest path problem on a specially crafted g raph, which enables us to enforce piecewise planarity. Besides producing compact piecewise planar models, this formulation allows us to di- rectly control the number of vertices (i.e., density) of the output mesh. We evaluate our system on real indoor scenes, and show that our texture mapped mesh models provide compelling free-viewpoint visualization experiences, when compared against the state-of-the-art and ground truth.

****

Data-driven Flower Petal Modeling with Botany Priors

Chenxi Zhang, Mao Ye, Bo Fu, Ruigang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 636-643

In this paper we focus on the 3D modeling of flower, in particular the petals. The complex structure, severe occlusions, and wide variations make the reconstruction of their 3D models a challenging task. Therefore, even though the flower is the most distinctive part of a plant, there has been little modeling study devoted to it. We overcome these challenges by combining data driven modeling techniques with domain knowledge from botany. Taking a 3D point cloud of an input flower scanned from a single view, our method starts with a level-set based segmentation of each individual petal, using both appearance and 3D information. Each segmented petal is then fitted with a scale-invariant morphable petal shape model, which is constructed from individually scanned exemplar petals. Novel constraints based on botany studies, such as the number and spatial layout of petals, are incorporated into the fitting process for realistically reconstructing occluded regions and maintaining correct 3D spatial relations. Finally, the reconstructed petal shape is texture mapped using the registered color images, with occluded regions filled in by content from visible ones. Experiments show that our approach can obtain realistic modeling of flowers even with severe occlusions and large shape/size variations.

****

User-Specific Hand Modeling from Monocular Depth Sequences

Jonathan Taylor, Richard Stebbing, Varun Ramakrishna, Cem Keskin, Jamie Shotton, Shahram Izadi, Aaron Hertzmann, Andrew Fitzgibbon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 644-651

This paper presents a method for acquiring dense nonrigid shape and deformation from a single monocular depth sensor. We focus on modeling the human hand, and assume that a single rough template model is available. We combine and extend existing work on model-based tracking, subdivision surface fitting, and mesh deformation to acquire detailed hand models from as few as 15 frames of depth data. We propose an objective that measures the error of fit between each sampled data point and a continuous model surface defined by a rigged control mesh, and uses as-rigid-as-possible (ARAP) regularizers to cleanly separate the model and template geometries. A key contribution is our use of a smooth model based on subdivision surfaces that allows simultaneous optimization over both correspondences and model parameters. This avoids the use of iterated closest point (ICP) algorithms which often lead to slow convergence. Automatic initialization is obtained using a regression forest trained to infer approximate correspondences. Experiments show that the resulting meshes model the user's hand shape more accurately than just adapting the shape parameters of the skeleton, and that the retargeted skeleton accurately models the user's articulations. We investigate the effect of various modeling choices, and show the benefits of using subdivision surfaces and ARAP regularization.

****

Class Specific 3D Object Shape Priors Using Surface Normals

Christian Hane, Nikolay Savinov, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 652-659

Dense 3D reconstruction of real world objects containing textureless, reflective and specular parts is a challenging task. Using general smoothness priors such as surface area regularization can lead to defects in the form of disconnected parts or unwanted indentations. We argue that this problem can be solved by exploiting the object class specific local surface orientations, e.g. a car is always close to horizontal in the roof area. Therefore, we formulate an object class s pecific shape prior in the form of spatially varying anisotropic smoothness terms. The parameters of the shape prior are extracted from training data. We detail how our shape prior formulation directly fits into recently proposed volumetric multi-label reconstruction approaches. This allows a segmentation between the object and its supporting ground. In our experimental evaluation we show reconstructions using our trained shape prior on several challenging datasets.

****

Frequency-Based 3D Reconstruction of Transparent and Specular Objects

Ding Liu, Xida Chen, Yee-Hong Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 660-667

3D reconstruction of transparent and specular objects is a very challenging topic in computer vision. For transparent and specular objects, which have complex interior and exterior structures that can reflect and refract light in a complex fashion, it is difficult, if not impossible, to use either passive stereo or the traditional structured light methods to do the reconstruction. We propose a frequency-based 3D reconstruction method, which incorporates the frequency-based matting method. Similar to the structured light methods, a set of frequency-based patterns are projected onto the object, and a camera captures the scene. Each pixel of the captured image is analyzed along the time axis and the corresponding signal is transformed to the frequency-domain using the Discrete Fourier Transform. Since the frequency is only determined by the source that creates it, the frequency of the signal can uniquely identify the location of the pixel in the patterns. In this way, the correspondences between the pixels in the captured images and the points in the patterns can be acquired. Using a new labelling procedure, the surface of transparent and specular objects can be reconstructed with very encouraging results.

****

Human Body Shape Estimation Using a Multi-Resolution Manifold Forest

Frank Perbet, Sam Johnson, Minh-Tri Pham, Bjorn Stenger; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 668-675

This paper proposes a method for estimating the 3D body shape of a person with robustness to clothing. We formulate the problem as optimization over the manifold of valid depth maps of body shapes learned from synthetic training data. The manifold itself is represented using a novel data structure, a Multi-Resolution Manifold Forest (MRMF), which contains vertical edges between tree nodes as well as horizontal edges between nodes across trees that correspond to overlapping partitions. We show that this data structure allows both efficient localization and navigation on the manifold for on-the-fly building of local linear models (manifold charting). We demonstrate shape estimation of clothed users, showing significant improvement in accuracy over global shape models and models using pre-computed clusters. We further compare the MRMF with alternative manifold charting methods on a public dataset for estimating 3D motion from noisy 2D marker observations, obtaining state-of-the-art results.

****

Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera

Qing Zhang, Bo Fu, Mao Ye, Ruigang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 676-683

In this paper we present a novel autonomous pipeline to build a personalized parametric model (pose-driven avatar) using a single depth sensor. Our method first captures a few high-quality scans of the user rotating herself at multiple poses from different views. We fit each incomplete scan using template fitting techniques with a generic human template, and register all scans to every pose using global consistency constraints. After registration, these watertight models with different poses are used to train a parametric model in a fashion similar to the SCAPE method. Once the parametric model is built, it can be used as an animitable avatar or more interestingly synthesizing dynamic 3D models from single-view depth videos. Experimental results demonstrate the effectiveness of our system to produce dynamic models.

****

Single-View 3D Scene Parsing by Attributed Grammar

Xiaobai Liu, Yibiao Zhao, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 684-691

In this paper, we present an attributed grammar for parsing man-made outdoor scenes into semantic surfaces, and recovering its 3D model simultaneously. The grammar takes superpixels as its terminal nodes and use five production rules to generate the scene into a hierarchical parse graph. Each graph node actually correlates with a surface or a composite of surfaces in the 3D world or the 2D image. They are described by attributes for the global scene model, e.g. focal length, vanishing points, or the surface properties, e.g. surface normal, contact line with other surfaces, and relative spatial location etc. Each production rule is associated with some equations that constraint the attributes of the parent nodes and those of their children nodes. Given an input image, our goal is to construct a hierarchical parse graph by recursively applying the five grammar rules while preserving the attributes constraints. We develop an effective top-down/bottom-up cluster sampling procedure which can explore this constrained space efficiently. We evaluate our method on both public benchmarks and newly built datasets, and achieve state-of-the-art performances in terms of layout estimation and region segmentation. We also demonstrate that our method is able to recover detailed 3D model with relaxed Manhattan structures which clearly advances the state-of-the-arts of singleview 3D reconstruction.

****

Separation of Line Drawings Based on Split Faces for 3D Object Reconstruction

Changqing Zou, Heng Yang, Jianzhuang Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 692-699

Reconstructing 3D objects from single line drawings is often desirable in computer vision and graphics applications. If the line drawing of a complex 3D object is decomposed into primitives of simple shape, the object can be easily reconstructed. We propose an effective method to conduct the line drawing separation and turn a complex line drawing into parametric 3D models. This is achieved by recursively separating the line drawing using two types of split faces. Our experiments show that the proposed separation method can generate more basic and simple line drawings, and its combination with the example-based reconstruction can robustly recover wider range of complex parametric 3D objects than previous methods

****

When 3D Reconstruction Meets Ubiquitous RGB-D Images

Quanshi Zhang, Xuan Song, Xiaowei Shao, Huijing Zhao, Ryosuke Shibasaki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 700-707

3D reconstruction from a single image is a classical problem in computer vision. However, it still poses great challenges for the reconstruction of daily-use objects with irregular shapes. In this paper, we propose to learn 3D reconstruction knowledge from informally captured RGB-D images, which will probably be ubiquitously used in daily life. The learning of 3D reconstruction is defined as a category modeling problem, in which a model for each category is trained to encode category-specific knowledge for 3D reconstruction. The category model estimates the pixel-level 3D structure of an object from its 2D appearance, by taking into account considerable variations in rotation, 3D structure, and texture. Learning 3D reconstruction from ubiquitous RGB-D images creates a new set of challenges. Experimental results have demonstrated the effectiveness of the proposed approach.

****

Stable Template-Based Isometric 3D Reconstruction in All Imaging Conditions by Linear Least-Squares

Ajad Chhatkuli, Daniel Pizarro, Adrien Bartoli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 708-715

It has been recently shown that reconstructing an isometric surface from a single 2D input image matched to a 3D template was a well-posed problem. This however does not tell us how reconstruction algorithms will behave in practical conditions, where the amount of perspective is generally small and the projection thus behaves like weak-perspective or orthography. We here bring answers to what is t heoretically recoverable in such imaging conditions, and explain why existing convex numerical solutions and analytical solutions to 3D reconstruction may be unstable. We then propose a new algorithm which works under all imaging conditions, from strong to loose perspective. We empirically show that the gain in stability is tremendous, bringing our results close to the iterative minimization of a statisticallyoptimal cost. Our algorithm has a low complexity, is simple and uses only one round of linear least-squares.

****

Discrete-Continuous Depth Estimation from a Single Image

Miaomiao Liu, Mathieu Salzmann, Xuming He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 716-723

In this paper, we tackle the problem of estimating the depth of a scene from a single image. This is a challenging task, since a single image on its own does not provide any depth cue. To address this, we exploit the availability of a pool of images for which the depth is known. More specifically, we formulate monocular depth estimation as a discrete-continuous optimization problem, where the continuous variables encode the depth of the superpixels in the input image, and the discrete ones represent relationships between neighboring superpixels. The solution to this discrete-continuous optimization problem is then obtained by performing inference in a graphical model using particle belief propagation. The unary potentials in this graphical model are computed by making use of the images with known depth. We demonstrate the effectiveness of our model in both the indoor and outdoor scenarios. Our experimental evaluation shows that our depth estimates are more accurate than existing methods on standard datasets.

****

Leveraging Hierarchical Parametric Networks for Skeletal Joints Based Action Segmentation and Recognition

Di Wu, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 724-731

Over the last few years, with the immense popularity of the Kinect, there has been renewed interest in developing methods for human gesture and action recognition from 3D skeletal data. A number of approaches have been proposed to extract representative features from 3D skeletal data, most commonly hard wired geometric or bio-inspired shape context features. We propose a hierarchial dynamic framework that first extracts high level skeletal joints features and then uses the learned representation for estimating emission probability to infer action sequences. Currently gaussian mixture models are the dominant technique for modeling the emission distribution of hidden Markov models. We show that better action recognition using skeletal features can be achieved by replacing gaussian mixture models by deep neural networks that contain many layers of features to predict probability distributions over states of hidden Markov models. The framework can be easily extended to include a ergodic state to segment and recognize actions simultaneously.

****

Seeing What You're Told: Sentence-Guided Activity Recognition In Video

Narayanaswamy Siddharth, Andrei Barbu, Jeffrey Mark Siskind; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 732-739

We present a system that demonstrates how the compositional structure of events, in concert with the compositional structure of language, can interplay with the underlying focusing mechanisms in video action recognition, providing a medium for top-down and bottom-up integration as well as multi-modal integration between vision and language. We show how the roles played by participants (nouns), their characteristics (adjectives), the actions performed (verbs), the manner of such actions (adverbs), and changing spatial relations between participants (prepositions), in the form of whole-sentence descriptions mediated by a grammar, guides the activity-recognition process. Further, the utility and expressiveness of our framework is demonstrated by performing three separate tasks in the domain of multi-activity video: sentence-guided focus of attention, generation of sentential description, and query-based search, simply by leveraging the framework in different manners.

****

Action Localization with Tubelets from Motion

Mihir Jain, Jan van Gemert, Herve Jegou, Patrick Bouthemy, Cees G.M. Snoek; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 740-747

This paper considers the problem of action localization, where the objective is to determine when and where certain actions appear. We introduce a sampling strategy to produce 2D+t sequences of bounding boxes, called tubelets. Compared to state-of-the-art alternatives, this drastically reduces the number of hypotheses that are likely to include the action of interest. Our method is inspired by a recent technique introduced in the context of image localization. Beyond considering this technique for the first time for videos, we revisit this strategy for 2D+t sequences obtained from super-voxels. Our sampling strategy advantageously exploits a criterion that reflects how action related motion deviates from background motion. We demonstrate the interest of our approach by extensive experiments on two public datasets: UCF Sports and MSR-II. Our approach significantly outperforms the state-of-the-art on both datasets, while restricting the search of actions to a fraction of possible bounding box sequences.

****

Actionness Ranking with Lattice Conditional Ordinal Random Fields

Wei Chen, Caiming Xiong, Ran Xu, Jason J. Corso; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 748-755

Action analysis in image and video has been attracting more and more attention in computer vision. Recognizing specific actions in video clips has been the main focus. We move in a new, more general direction in this paper and ask the critical fundamental question: what is action, how is action different from motion, and in a given image or video where is the action? We study the philosophical and visual characteristics of action, which lead us to define actionness: intentional bodily movement of biological agents (people, animals). To solve the general problem, we propose the lattice conditional ordinal random field model that incorporates local evidence as well as neighboring order agreement. We implement the new model in the continuous domain and apply it to scoring actionness in both image and video datasets. Our experiments demonstrate not only that our new model can outperform the popular ranking SVM but also that indeed action is distinct from motion.

****

Multiple Granularity Analysis for Fine-grained Action Detection

Bingbing Ni, Vignesh R. Paramathayalan, Pierre Moulin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 756-763

We propose to decompose the fine-grained human activity analysis problem into two sequential tasks with increasing granularity. Firstly, we infer the coarse interaction status, i.e., which object is being manipulated and where it is. Knowing that the major challenge is frequent mutual occlusions during manipulation, we propose an "interaction tracking" framework in which hand/object position and interaction status are jointly tracked by explicitly modeling the contextual information between mutual occlusion and interaction status. Secondly, the inferred hand/object position and interaction status are utilized to provide 1) more compact feature pooling by effectively pruning large number of motion features from irrelevant spatio-temporal positions and 2) discriminative action detection by a granularity fusion strategy. Comprehensive experiments on two challenging fine-grained activity datasets (i.e., cooking action) show that the proposed framework achieves high accuracy/robustness in tracking multiple mutually occluded hands/objects during manipulation as well as significant performance improvement on fine-grained action detection over state-of-the-art methods.

****

Human Action Recognition Across Datasets by Foreground-weighted Histogram Decomposition

Waqas Sultani, Imran Saleemi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 764-771

This paper attempts to address the problem of recognizing human actions while training and testing on distinct datasets, when test videos are neither labeled nor available during training. In this scenario, learning of a joint vocabulary, or domain transfer techniques are not applicable. We first explore reasons for poor classifier performance when tested on novel datasets, and quantify the effect of scene backgrounds on action representations and recognition. Using only the background features and partitioning of gist feature space, we show that the background scenes in recent datasets are quite discriminative and can be used classify an action with reasonable accuracy. We then propose a new process to obtain a measure of confidence in each pixel of the video being a foreground region, using motion, appearance, and saliency together in a 3D MRF based framework. We also propose multiple ways to exploit the foreground confidence: to improve bag-of-words vocabulary, histogram representation of a video, and a novel histogram decomposition based representation and kernel. We used these foreground confidences to recognize actions trained on one data set and test on a different data set. We have performed extensive experiments on several datasets that improve cross dataset recognition accuracy as compared to baseline methods.
****

Range-Sample Depth Feature for Action Recognition
Cewu Lu, Jiaya Jia, Chi-Keung Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 772-779
We propose binary range-sample feature in depth. It is based on t tests and achieves reasonable invariance with respect to possible change in scale, viewpoint, and background. It is robust to occlusion and data corruption as well. The descriptor works in a high speed thanks to its binary property. Working together with standard learning algorithms, the proposed descriptor achieves state-of-theart results on benchmark datasets in our experiments. Impressively short running time is also yielded.
****

The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities
Hilde Kuehne, Ali Arslan, Thomas Serre; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 780-787
This paper describes a framework for modeling human activities as temporally structured processes. Our approach is motivated by the inherently hierarchical nature of human activities and the close correspondence between human actions and speech: We model action units using Hidden Markov Models, much like words in speech. These action units then form the building blocks to model complex human activities as sentences using an action grammar. To evaluate our approach, we collected a large dataset of daily cooking activities: The dataset includes a total of 52 participants, each performing a total of 10 cooking activities in multiple real-life kitchens, resulting in over 77 hours of video footage. We evaluate the HTK toolkit, a state-of-the-art speech recognition engine, in combination with multiple video feature descriptors, for both the recognition of cooking activities (e.g., making pancakes) as well as the semantic parsing of videos into action units (e.g., cracking eggs). Our results demonstrate the benefits of structured temporal generative approaches over existing discriminative approaches in coping with the complexity of human daily life activities.
****

Complex Activity Recognition using Granger Constrained DBN (GCDBN) in Sports and Surveillance Video
Eran Swears, Anthony Hoogs, Qiang Ji, Kim Boyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 788-795
Modeling interactions of multiple co-occurring objects in a complex activity is becoming increasingly popular in the video domain. The Dynamic Bayesian Network (DBN) has been applied to this problem in the past due to its natural ability to statistically capture complex temporal dependencies. However, standard DBN structure learning algorithms are generatively learned, require manual structure definitions, and/or are computationally complex or restrictive. We propose a novel structure learning solution that fuses the Granger Causality statistic, a direct

measure of temporal dependence, with the Adaboost feature selection algorithm to automatically constrain the temporal links of a DBN in a discriminative manner. This approach enables us to completely define the DBN structure prior to parameter learning, which reduces computational complexity in addition to providing a more descriptive structure. We refer to this modeling approach as the Granger Constraints DBN (GCDBN). Our experiments show how the GCDBN outperforms two of the most relevant state-of-the-art graphical models in complex activity classification on handball video data, surveillance data, and synthetic data.
****
Incremental Activity Modeling and Recognition in Streaming Videos
Mahmudul Hasan, Amit K. Roy-Chowdhury; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 796-803
Most of the state-of-the-art approaches to human activity recognition in video need an intensive training stage and assume that all of the training examples are labeled and available beforehand. But these assumptions are unrealistic for many applications where we have to deal with streaming videos. In these videos, as new activities are seen, they can be leveraged upon to improve the current activity recognition models. In this work, we develop an incremental activity learning framework that is able to continuously update the activity models and learn new ones as more videos are seen. Our proposed approach leverages upon state-of-the-art machine learning tools, most notably active learning systems. It does not require tedious manual labeling of every incoming example of each activity class. We perform rigorous experiments on challenging human activity datasets, which demonstrate that the incremental activity modeling framework can achieve performance very close to the cases when all examples are available a priori.
****
Super Normal Vector for Activity Recognition Using Depth Sequences
Xiaodong Yang, YingLi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 804-811
This paper presents a new framework for human activity recognition from video sequences captured by a depth camera. We cluster hypersurface normals in a depth sequence to form the polynormal which is used to jointly characterize the local motion and shape information. In order to globally capture the spatial and temporal orders, an adaptive spatio-temporal pyramid is introduced to subdivide a depth video into a set of space-time grids. We then propose a novel scheme of aggregating the low-level polynormals into the super normal vector (SNV) which can be seen as a simplified version of the Fisher kernel representation. In the extensive experiments, we achieve classification results superior to all previous published results on the four public benchmark datasets, i.e., MSRAction3D, MSRDailyActivity3D, MSRGesture3D, and MSRActionPairs3D.
****
Discriminative Hierarchical Modeling of Spatio-Temporally Composable Human Activities
Ivan Lillo, Alvaro Soto, Juan Carlos Niebles; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 812-819
This paper proposes a framework for recognizing complex human activities in videos. Our method describes human activities in a hierarchical discriminative model that operates at three semantic levels. At the lower level, body poses are encoded in a representative but discriminative pose dictionary. At the intermediate level, encoded poses span a space where simple human actions are composed. At the highest level, our model captures temporal and spatial compositions of actions into complex human activities. Our human activity classifier simultaneously models which body parts are relevant to the action of interest as well as their appearance and composition using a discriminative approach. By formulating model learning in a max-margin framework, our approach achieves powerful multi-class discrimination while providing useful annotations at the intermediate semantic level. We show how our hierarchical compositional model provides natural handling of occlusions. To evaluate the effectiveness of our proposed framework, we introduce a new dataset of composed human activities. We provide empirical evidence that our method achieves state-of-the-art activity classification performance on se

veral benchmark datasets.

****

A Multigraph Representation for Improved Unsupervised/Semi-supervised Learning of Human Actions

Simon Jones, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 820-826

Graph-based methods are a useful class of methods for improving the performance of unsupervised and semi-supervised machine learning tasks, such as clustering or information retrieval. However, the performance of existing graph-based methods is highly dependent on how well the affinity graph reflects the original data structure. We propose that multimedia such as images or videos consist of multiple separate components, and therefore more than one graph is required to fully capture the relationship between them. Accordingly, we present a new spectral method - the Feature Grouped Spectral Multigraph (FGSM) - which comprises the following steps. First, mutually independent subsets of the original feature space are generated through feature clustering. Secondly, a separate graph is generated from each feature subset. Finally, a spectral embedding is calculated on each graph, and the embeddings are scaled/aggregated into a single representation. Using this representation, a variety of experiments are performed on three learning tasks - clustering, retrieval and recognition - on human action datasets, demonstrating considerably better performance than the state-of-the-art.

****

StoryGraphs: Visualizing Character Interactions as a Timeline

Makarand Tapaswi, Martin Bauml, Rainer Stiefelhagen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 827-834

We present a novel way to automatically summarize and represent the storyline of a TV episode by visualizing character interactions as a chart. We also propose a scene detection method that lends itself well to generate over-segmented scenes which is used to partition the video. The positioning of character lines in the chart is formulated as an optimization problem which trades between the aesthetics and functionality of the chart. Using automatic person identification, we present StoryGraphs for 3 diverse TV series encompassing a total of 22 episodes. We define quantitative criteria to evaluate StoryGraphs and also compare them against episode summaries to evaluate their ability to provide an overview of the episode.

****

Learning Receptive Fields for Pooling from Tensors of Feature Response

Can Xu, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 835-842

A new method for learning pooling receptive fields for recognition is presented. The method exploits the statistics of the 3D tensor of SIFT responses to an image. It is argued that the eigentensors of this tensor contain the information necessary for learning class-specific pooling recep- tive fields. It is shown that this information can be extracted by a simple PCA analysis of a specific tensor flattening. A novel algorithm is then proposed for fitting box-like receptive fields to the eigenimages extracted from a collection of images. The resulting receptive fields can be combined with any of the recently popular coding strategies for image classification. This combination is experimentally shown to improve classification accuracy for both vector quantization and Fisher vector (FV) encodings. It is then shown that the combination of the FV encoding with the proposed receptive fields has state-of-the-art performance for both object recognition and scene classification. Finally, when compared with previous attempts at learning receptive fields for pooling, the method is simpler and achieves better results.

****

Towards Unified Human Parsing and Pose Estimation

Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 843-850

We study the problem of human body configuration analysis, more specifically, hu man parsing and human pose estimation. These two tasks, i.e. identifying the semantic regions and body joints respectively over the human body image, are intrinsically highly correlated. However, previous works generally solve these two problems separately or iteratively. In this work, we propose a unified framework for simultaneous human parsing and pose estimation based on semantic parts. By utilizing Parselets and Mixture of Joint-Group Templates as the representations for these semantic parts, we seamlessly formulate the human parsing and pose estimation problem jointly within a unified framework via a tailored And-Or graph. A novel Grid Layout Feature is then designed to effectively capture the spatial co-occurrence/occlusion information between/within the Parselets and MJGTs. Thus the mutually complementary nature of these two tasks can be harnessed to boost the performance of each other.The resultant unified model can be solved using the structure learning framework in a principled way. Comprehensive evaluations on two benchmark datasets for both tasks demonstrate the effectiveness of the proposed framework when compared with the state-of-the-art methods.

****

Ask the Image: Supervised Pooling to Preserve Feature Locality

Sean Ryan Fanello, Nicoletta Noceti, Carlo Ciliberto, Giorgio Metta, Francesca Odone; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 851-858

In this paper we propose a weighted supervised pooling method for visual recognition systems. We combine a standard Spatial Pyramid Representation which is commonly adopted to encode spatial information, with an appropriate Feature Space Representation favoring semantic information in an appropriate feature space. For the latter, we propose a weighted pooling strategy exploiting data supervision to weigh each local descriptor coherently with its likelihood to belong to a given object class. The two representations are then combined adaptively with Multiple Kernel Learning. Experiments on common benchmarks (Caltech-256 and PASCAL VOC-2007) show that our image representation improves the current visual recognition pipeline and it is competitive with similar state-of-art pooling methods. We also evaluate our method on a real Human-Robot Interaction setting, where the pure Spatial Pyramid Representation does not provide sufficient discriminative power, obtaining a remarkable improvement.

****

Similarity Comparisons for Interactive Fine-Grained Categorization

Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, Serge Belongie; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 859-866

Current human-in-the-loop fine-grained visual categorization systems depend on a predefined vocabulary of attributes and parts, usually determined by experts. In this work, we move away from that expert-driven and attribute-centric paradigm and present a novel interactive classification system that incorporates computer vision and perceptual similarity metrics in a unified framework. At test time, users are asked to judge relative similarity between a query image and various sets of images; these general queries do not require expert-defined terminology and are applicable to other domains and basic-level categories, enabling a flexible, efficient, and scalable system for fine-grained categorization with humans in the loop. Our system outperforms existing state-of-the-art systems for relevance feedback-based image retrieval as well as interactive classification, resulting in a reduction of up to 43% in the average number of questions needed to correctly classify an image.

****

Continuous Manifold Based Adaptation for Evolving Visual Domains

Judy Hoffman, Trevor Darrell, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 867-874

We pose the following question: what happens when test data not only differs from training data, but differs from it in a continually evolving way? The classic domain adaptation paradigm considers the world to be separated into stationary domains with clear boundaries between them. However, in many real-world applications, examples cannot be naturally separated into discrete domains, but arise fro m a continuously evolving underlying process. Examples include video with gradually changing lighting and spam email with evolving spammer tactics. We formulate a novel problem of adapting to such continuous domains, and present a solution based on smoothly varying embeddings. Recent work has shown the utility of considering discrete visual domains as fixed points embedded in a manifold of lower-dimensional subspaces. Adaptation can be achieved via transforms or kernels learned between such stationary source and target subspaces. We propose a method to consider non-stationary domains, which we refer to as Continuous Manifold Adaptation (CMA). We treat each target sample as potentially being drawn from a different subspace on the domain manifold, and present a novel technique for continuous transform-based adaptation. Our approach can learn to distinguish categories using training data collected at some point in the past, and continue to update its model of the categories for some time into the future, without receiving any additional labels. Experiments on two visual datasets demonstrate the value of our approach for several popular feature representations.

****

Talking Heads: Detecting Humans and Recognizing Their Interactions

Minh Hoai, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 875-882

The objective of this work is to accurately and efficiently detect configurations of one or more people in edited TV material. Such configurations often appear in standard arrangements due to cinematic style, and we take advantage of this to provide scene context. We make the following contributions: first, we introduce a new learnable context aware configuration model for detecting sets of people in TV material that predicts the scale and location of each upper body in the configuration; second, we show that inference of the model can be solved globally and efficiently using dynamic programming, and implement a maximum margin learning framework; and third, we show that the configuration model substantially outperforms a Deformable Part Model (DPM) for predicting upper body locations in video frames, even when the DPM is equipped with the context of other upper bodies. Experiments are performed over two datasets: the TV Human Interaction dataset, and 150 episodes from four different TV shows. We also demonstrate the benefits of the model in recognizing interactions in TV shows.

****

Salient Region Detection via High-Dimensional Color Transform

Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 883-890

In this paper, we introduce a novel technique to automatically detect salient regions of an image via high-dimensional color transform. Our main idea is to represent a saliency map of an image as a linear combination of high-dimensional color space where salient regions and backgrounds can be distinctively separated. This is based on an observation that salient regions often have distinctive colors compared to the background in human perception, but human perception is often complicated and highly nonlinear. By mapping a low dimensional RGB color to a feature vector in a high-dimensional color space, we show that we can linearly separate the salient regions from the background by finding an optimal linear combination of color coefficients in the high-dimensional color space. Our high dimensional color space incorporates multiple color representations including RGB, CIELab, HSV and with gamma corrections to enrich its representative power. Our experimental results on three benchmark datasets show that our technique is effective, and it is computationally efficient in comparison to previous state-of-the-art techniques.

****

The Role of Context for Object Detection and Semantic Segmentation in the Wild

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, Alan Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 891-898

In this paper we study the role of context in existing state-of-the-art detection and segmentation approaches. Towards this goal, we label every pixel of PASCAL VOC 2010 detection challenge with a semantic category. We believe this data wil l provide plenty of challenges to the community, as it contains 520 additional classes for semantic segmentation and object detection. Our analysis shows that nearest neighbor based approaches perform poorly on semantic segmentation of contextual classes, showing the variability of PASCAL imagery. Furthermore, improvements of exist ing contextual models for detection is rather modest. In order to push forward the performance in this difficult scenario, we propose a novel deformable part-based model, which exploits both local context around each candidate detection as well as global context at the level of the scene. We show that this contextual reasoning significantly helps in detecting objects at all scales.

****

Switchable Deep Network for Pedestrian Detection

Ping Luo, Yonglong Tian, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 899-906

In this paper, we propose a Switchable Deep Network (SDN) for pedestrian detection. The SDN automatically learns hierarchical features, salience maps, and mixture representations of different body parts. Pedestrian detection faces the challenges of background clutter and large variations of pedestrian appearance due to pose and viewpoint changes and other factors. One of our key contributions is to propose a Switchable Restricted Boltzmann Machine (SRBM) to explicitly model the complex mixture of visual variations at multiple levels. At the feature levels, it automatically estimates saliency maps for each test sample in order to separate background clutters from discriminative regions for pedestrian detection. At the part and body levels, it is able to infer the most appropriate template for the mixture models of each part and the whole body. We have devised a new generative algorithm to effectively pretrain the SDN and then fine-tune it with back-propagation. Our approach is evaluated on the Caltech and ETH datasets and achieves the state-of-the-art detection performance.

****

Compact Representation for Image Classification: To Choose or to Compress?

Yu Zhang, Jianxin Wu, Jianfei Cai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 907-914

In large scale image classification, features such as Fisher vector or VLAD have achieved state-of-the-art results. However, the combination of large number of examples and high dimensional vectors necessitates dimensionality reduction, in order to reduce its storage and CPU costs to a reasonable range. In spite of the popularity of various feature compression methods, this paper argues that feature selection is a better choice than feature compression. We show that strong multicollinearity among feature dimensions may not exist, which undermines feature compression's effectiveness and renders feature selection a natural choice. We also show that many dimensions are noise and throwing them away is helpful for classification. We propose a supervised mutual information (MI) based importance sorting algorithm to choose features. Combining with 1-bit quantization, MI feature selection has achieved both higher accuracy and less computational cost than feature compression methods such as product quantization and BPBC.

****

Capturing Long-tail Distributions of Object Subcategories

Xiangxin Zhu, Dragomir Anguelov, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 915-922

We argue that object subcategories follow a long-tail distribution: a few subcategories are common, while many are rare. We describe distributed algorithms for learning large- mixture models that capture long-tail distributions, which are hard to model with current approaches. We introduce a generalized notion of mixtures (or subcategories) that allow for examples to be shared across multiple subcategories. We optimize our models with a discriminative clustering algorithm that searches over mixtures in a distributed, "brute-force" fashion. We used our scalable system to train tens of thousands of deformable mixtures for VOC objects. We demonstrate significant performance improvements, particularly for object classes that are characterized by large appearance variation.

****

Accurate Object Detection with Joint Classification-Regression Random Forests

Samuel Schulter, Christian Leistner, Paul Wohlhart, Peter M. Roth, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 923-930

In this paper, we present a novel object detection approach that is capable of regressing the aspect ratio of objects. This results in accurately predicted bounding boxes having high overlap with the ground truth. In contrast to most recent works, we employ a Random Forest for learning a template-based model but exploit the nature of this learning algorithm to predict arbitrary output spaces. In this way, we can simultaneously predict the object probability of a window in a sliding window approach as well as regress its aspect ratio with a single model. Furthermore, we also exploit the additional information of the aspect ratio during the training of the Joint Classification-Regression Random Forest, resulting in better detection models.   Our experiments demonstrate several benefits: (i) Our approach gives competitive results on standard detection benchmarks. (ii) The additional aspect ratio regression delivers more accurate bounding boxes than standard object detection approaches in terms of overlap with ground truth, especially when tightening the evaluation criterion. (iii) The detector itself becomes better by only including the aspect ratio information during training.

****

Additive Quantization for Extreme Vector Compression

Artem Babenko, Victor Lempitsky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 931-938

We introduce a new compression scheme for high-dimensional vectors that approximates the vectors using sums of M codewords coming from M different codebooks. We show that the proposed scheme permits efficient distance and scalar product computations between compressed and uncompressed vectors. We further suggest vector encoding and codebook learning algorithms that can minimize the coding error within the proposed scheme. In the experiments, we demonstrate that the proposed compression can be used instead of or together with product quantization. Compared to product quantization and its optimized versions, the proposed compression approach leads to lower coding approximation errors, higher accuracy of approximate nearest neighbor search in the datasets of visual descriptors, and lower image classification error, whenever the classifiers are learned on or applied to compressed vectors.

****

Product Sparse Coding

Tiezheng Ge, Kaiming He, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 939-946

Sparse coding is a widely involved technique in computer vision. However, the expensive computational cost can hamper its applications, typically when the codebook size must be limited due to concerns on running time. In this paper, we study a special case of sparse coding in which the codebook is a Cartesian product of two subcodebooks. We present algorithms to decompose this sparse coding problem into smaller subproblems, which can be separately solved. Our solution, named as Product Sparse Coding (PSC), reduces the time complexity from O(K) to O(sqrt(K)) in the codebook size K. In practice, this can be 20-100x faster than standard sparse coding. In experiments we demonstrate the efficiency and quality of this method on the applications of image classification and image retrieval.

****

Informed Haar-like Features Improve Pedestrian Detection

Shanshan Zhang, Christian Bauckhage, Armin B. Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 947-954

We propose a simple yet effective detector for pedestrian detection. The basic idea is to incorporate common sense and everyday knowledge into the design of simple and computationally efficient features. As pedestrians usually appear up-right in image or video data, the problem of pedestrian detection is considerably simpler than general purpose people detection. We therefore employ a statistical model of the up-right human body where the head, the upper body, and the lower body are treated as three distinct components. Our main contribution is to systematically design a pool of rectangular templates that are tailored to this shape model. As we incorporate different kinds of low-level measurements, the resulting multi-modal & multi-channel Haar-like features represent characteristic differences between parts of the human body yet are robust against variations in clothing or environmental settings. Our approach avoids exhaustive searches over all possible configurations of rectangle features and neither relies on random sampling. It thus marks a middle ground among recently published techniques and yields efficient low-dimensional yet highly discriminative features. Experimental results on the INRIA and Caltech pedestrian datasets show that our detector reaches state-of-the-art performance at low computational costs and that our features are robust against occlusions.

****

Image Reconstruction from Bag-of-Visual-Words

Hiroharu Kato, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 955-962

The objective of this study is to reconstruct images from Bag-of-Visual-Words (BoVW), which is the de facto standard feature for image retrieval and recognition. BoVW is defined here as a histogram of quantized descriptors extracted densely on a regular grid at a single scale. Despite its wide use, no report describes reconstruction of the original image of a BoVW. This task is challenging for two reasons: 1) BoVW includes quantization errors when local descriptors are assigned to visual words. 2) BoVW lacks spatial information of local descriptors when we count the occurrence of visual words. To tackle this difficult task, we use a large-scale image database to estimate the spatial arrangement of local descriptors. Then this task creates a jigsaw puzzle problem with adjacency and global location costs of visual words. Solving this optimization problem is also challenging because it is known as an NP-Hard problem. We propose a heuristic but efficient method to optimize it. To underscore the effectiveness of our method, we apply it to BoVWs extracted from about 100 different categories and demonstrate that it can reconstruct the original images, although the image features lack spatial information and include quantization errors.

****

Beta Process Multiple Kernel Learning

Bingbing Ni, Teng Li, Pierre Moulin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 963-970

In kernel based learning, the kernel trick transforms the original representation of a feature instance into a vector of similarities with the training feature instances, known as kernel representation. However, feature instances are sometimes ambiguous and the kernel representation calculated based on them do not possess any discriminative information, which can eventually harm the trained classifier. To address this issue, we propose to automatically select good feature instances when calculating the kernel representation in multiple kernel learning. Specifically, for the kernel representation calculated for each input feature instance, we multiply it element-wise with a latent binary vector named as instance selection variables, which targets at selecting good instances and attenuate the effect of ambiguous ones in the resulting new kernel representation. Beta process is employed for generating the prior distribution for the latent instance selection variables. We then propose a Bayesian graphical model which integrates both MKL learning and inference for the distribution of the latent instance selection variables. Variational inference is derived for model learning under a max-margin principle. Our method is called Beta process multiple kernel learning. Extensive experiments demonstrate the effectiveness of our method on instance selection and its high discriminative capability for various classification problems in vision.

****

Random Laplace Feature Maps for Semigroup Kernels on Histograms

Jiyan Yang, Vikas Sindhwani, Quanfu Fan, Haim Avron, Michael W. Mahoney; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 971-978

With the goal of accelerating the training and testing complexity of nonlinear kernel methods, several recent papers have proposed explicit embeddings of the in put data into low-dimensional feature spaces, where fast linear methods can instead be used to generate approximate solutions. Analogous to random Fourier feature maps to approximate shift-invariant kernels, such as the Gaussian kernel, we develop a new randomized technique called random Laplace features, to approximate a family of kernel functions adapted to the semigroup structure. This is the natural algebraic structure on the set of histograms and other non-negative data representations. We provide theoretical results on the uniform convergence of random Laplace features. Empirical analyses on image classification and surveillance event detection tasks demonstrate the attractiveness of using random Laplace features relative to several other feature maps proposed in the literature.

****

Hash-SVM: Scalable Kernel Machines for Large-Scale Visual Classification

Yadong Mu, Gang Hua, Wei Fan, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 979-986

This paper presents a novel algorithm which uses compact hash bits to greatly improve the efficiency of non-linear kernel SVM in very large scale visual classification problems. Our key idea is to represent each sample with compact hash bits, over which an inner product is defined to serve as the surrogate of the original nonlinear kernels. Then the problem of solving the nonlinear SVM can be transformed into solving a linear SVM over the hash bits. The proposed Hash-SVM enjoys dramatic storage cost reduction owing to the compact binary representation, as well as a (sub-)linear training complexity via linear SVM. As a critical component of Hash-SVM, we propose a novel hashing scheme for arbitrary non-linear kernels via random subspace projection in reproducing kernel Hilbert space. Our comprehensive analysis reveals a well behaved theoretic bound of the deviation between the proposed hashing-based kernel approximation and the original kernel function. We also derive requirements on the hash bits for achieving a satisfactory accuracy level. Several experiments on large-scale visual classification benchmarks are conducted, including one with over 1 million images. The results show that Hash-SVM greatly reduces the computational complexity (more than ten times faster in many cases) while keeping comparable accuracies.

****

Transitive Distance Clustering with K-Means Duality

Zhiding Yu, Chunjing Xu, Deyu Meng, Zhuo Hui, Fanyi Xiao, Wenbo Liu, Jianzhuang Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 987-994

We propose a very intuitive and simple approximation for the conventional spectral clustering methods. It effectively alleviates the computational burden of spectral clustering - reducing the time complexity from O(n^3) to O(n^2) - while capable of gaining better performance in our experiments. Specifically, by involving a more realistic and effective distance and the "k-means duality" property, our algorithm can handle datasets with complex cluster shapes, multi-scale clusters and noise. We also show its superiority in a series of its real applications on tasks including digit clustering as well as image segmentation.

****

Simultaneous Twin Kernel Learning using Polynomial Transformations for Structured Prediction

Chetan Tonde, Ahmed Elgammal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 995-1002

Many learning problems in computer vision can be posed as structured prediction problems, where the input and output instances are structured objects such as trees, graphs or strings rather than, single labels {+1, -1} or scalars. Kernel methods such as Structured Support Vector Machines , Twin Gaussian Processes (TGP), Structured Gaussian Processes, and vector-valued Reproducing Kernel Hilbert Spaces (RKHS), offer powerful ways to perform learning and inference over these domains. Positive definite kernel functions allow us to quantitatively capture similarity between a pair of instances over these arbitrary domains. A poor choice of the kernel function, which decides the RKHS feature space, often results in poor performance. Automatic kernel selection methods have been developed, but have focused only on kernels on the input domain (i.e.'one-way'). In this work, we propose a novel and efficient algorithm for learning kernel functions simultaneously, on both input and output domains. We introduce the idea of learning polynomial kernel transformations, and call this method Simultaneous Twin Kernel Learning (STKL). STKL can learn arbitrary, but continuous kernel functions, including 'one-way' kernel learning as a special case. We formulate this problem for learning covariances kernels of Twin Gaussian Processes. Our experimental evaluation using learned kernels on synthetic and several real-world datasets demonstrate consistent improvement in performance of TGP's.

****

Bregman Divergences for Infinite Dimensional Covariance Matrices

Mehrtash Harandi, Mathieu Salzmann, Fatih Porikli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1003-1010

We introduce an approach to computing and comparing Covariance Descriptors (CovDs) in infinite-dimensional spaces. CovDs have become increasingly popular to address classification problems in computer vision. While CovDs offer some robustness to measurement variations, they also throw away part of the information contained in the original data by only retaining the second-order statistics over the measurements. Here, we propose to overcome this limitation by first mapping the original data to a high-dimensional Hilbert space, and only then compute the CovDs. We show that several Bregman divergences can be computed between the resulting CovDs in Hilbert space via the use of kernels. We then exploit these divergences for classification purpose. Our experiments demonstrate the benefits of our approach on several tasks, such as material and texture recognition, person re-identification, and action recognition from motion capture data.

****

Optimizing Average Precision using Weakly Supervised Data

Aseem Behl, C. V. Jawahar, M. Pawan Kumar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1011-1018

The performance of binary classification tasks, such as action classification and object detection, is often measured in terms of the average precision (AP). Yet it is common practice in computer vision to employ the support vector machine (SVM) classifier, which optimizes a surrogate 0-1 loss. The popularity of SVM can be attributed to its empirical performance. Specifically, in fully supervised settings, SVM tends to provide similar accuracy to the AP-SVM classifier, which directly optimizes an AP-based loss. However, we hypothesize that in the significantly more challenging and practically useful setting of weakly supervised learning, it becomes crucial to optimize the right accuracy measure. In order to test this hypothesis, we propose a novel latent AP-SVM that minimizes a carefully designed upper bound on the AP-based loss function over weakly supervised samples. Using publicly available datasets, we demonstrate the advantage of our approach over standard loss-based binary classifiers on two challenging problems: action classification and character recognition.

****

Subspace Clustering for Sequential Data

Stephen Tierney, Junbin Gao, Yi Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1019-1026

We propose Ordered Subspace Clustering (OSC) to segment data drawn from a sequentially ordered union of subspaces. Current subspace clustering techniques learn the relationships within a set of data and then use a separate clustering algorithm such as NCut for final segmentation. In contrast our technique, under certain conditions, is capable of segmenting clusters intrinsically without providing the number of clusters as a parameter. Similar to Sparse Subspace Clustering (SSC) we formulate the problem as one of finding a sparse representation but include a new penalty term to take care of sequential data. We test our method on data drawn from infrared hyper spectral data, video sequences and face images. Our experiments show that our method, OSC, outperforms the state of the art methods: Spatial Subspace Clustering (SpatSC), Low-Rank Representation (LRR) and SSC.

****

Predicting Multiple Attributes via Relative Multi-task Learning

Lin Chen, Qiang Zhang, Baoxin Li; Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), 2014, pp. 1027-1034

Relative attributes learning aims to learn ranking functions describing the relative strength of attributes. Most of current learning approaches learn ranking functions for each attribute independently without considering possible intrinsic relatedness among the attributes. For a problem involving multiple attributes, it is reasonable to assume that utilizing such relatedness among the attributes would benefit learning, especially when the number of labeled training pairs are very limited. In this paper, we proposed a relative multi-attribute learning framework that integrates relative attributes into a multi-task learning scheme. The formulation allows us to exploit the advantages of the state-of-the-art regularization-based multi-task learning for improved attribute learning. In particular, using joint feature learning as the case studies, we evaluated our framework with both synthetic data and two real datasets. Experimental results suggest that the proposed framework has clear performance gain in ranking accuracy and zero-shot learning accuracy over existing methods of independent relative attributes learning and multi-task learning.

****

Learning Inhomogeneous FRAME Models for Object Patterns

Jianwen Xie, Wenze Hu, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1035-1042

We investigate an inhomogeneous version of the FRAME (Filters, Random field, And Maximum Entropy) model and apply it to modeling object patterns. The inhomogeneous FRAME is a non-stationary Markov random field model that reproduces the observed marginal distributions or statistics of filter responses at all the different locations, scales and orientations. Our experiments show that the inhomogeneous FRAME model is capable of generating a wide variety of object patterns in natural images. We then propose a sparsified version of the inhomogeneous FRAME model where the model reproduces observed statistical properties of filter responses at a small number of selected locations, scales and orientations. We propose to select these locations, scales and orientations by a shared sparse coding scheme, and we explore the connection between the sparse FRAME model and the linear additive sparse coding model. Our experiments show that it is possible to learn sparse FRAME models in unsupervised fashion and the learned models are useful for object classification.

****

Empirical Minimum Bayes Risk Prediction: How to Extract an Extra Few % Performance from Vision Models with Just Three More Parameters

Vittal Premachandran, Daniel Tarlow, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1043-1050

When building vision systems that predict structured objects such as image segmentations or human poses, a crucial concern is performance under task-specific evaluation measures (e.g. Jaccard Index or Average Precision). An ongoing research challenge is to optimize predictions so as to maximize performance on such complex measures. In this work, we present a simple meta-algorithm that is surprisingly effective â■■ Empirical Min Bayes Risk. EMBR takes as input a pre-trained model that would normally be the final product and learns three additional parameters so as to optimize performance on the complex high-order task-specific measure. We demonstrate EMBR in several domains, taking existing state-of-the-art algorithms and improving performance up to ~7%, simply with three extra parameters.

****

Fantope Regularization in Metric Learning

Marc T. Law, Nicolas Thome, Matthieu Cord; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1051-1058

This paper introduces a regularization method to explicitly control the rank of a learned symmetric positive semidefinite distance matrix in distance metric learning. To this end, we propose to incorporate in the objective function a linear regularization term that minimizes the k smallest eigenvalues of the distance matrix. It is equivalent to minimizing the trace of the product of the distance matrix with a matrix in the convex hull of rank-k projection matrices, called a Fantope. Based on this new regularization method, we derive an optimization schem e to efficiently learn the distance matrix. We demonstrate the effectiveness of the method on synthetic and challenging real datasets of face verification and image classification with relative attributes, on which our method outperforms state-of-the-art metric learning algorithms.

****

Kernel-PCA Analysis of Surface Normals for Shape-from-Shading

Patrick Snape, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1059-1066

We propose a kernel-based framework for computing components from a set of surface normals. This framework allows us to easily demonstrate that component analysis can be performed directly upon normals. We link previously proposed mapping functions, the azimuthal equidistant projection (AEP) and principal geodesic analysis (PGA), to our kernel-based framework. We also propose a new mapping function based upon the cosine distance between normals. We demonstrate the robustness of our proposed kernel when trained with noisy training sets. We also compare our kernels within an existing shape-from-shading (SFS) algorithm. Our spherical representation of normals, when combined with the robust properties of cosine kernel, produces a very robust subspace analysis technique. In particular, our results within SFS show a substantial qualitative and quantitative improvement over existing techniques.

****

Merging SVMs with Linear Discriminant Analysis: A Combined Model

Symeon Nikitidis, Stefanos Zafeiriou, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1067-1074

A key problem often encountered by many learning algorithms in computer vision dealing with high dimensional data is the so called "curse of dimensionality" which arises when the available training samples are less than the input feature space dimensionality. To remedy this problem, we propose a joint dimensionality reduction and classification framework by formulating an optimization problem within the maximum margin class separation task. The proposed optimization problem is solved using alternative optimization where we jointly compute the low dimensional maximum margin projections and the separating hyperplanes in the projection subspace. Moreover, in order to reduce the computational cost of the developed optimization algorithm we incorporate orthogonality constraints on the derived projection bases and show that the resulting combined model is an alternation between identifying the optimal separating hyperplanes and performing a linear discriminant analysis on the support vectors. Experiments on face, facial expression and object recognition validate the effectiveness of the proposed method against state-of-the-art dimensionality reduction algorithms.

****

Stable Learning in Coding Space for Multi-Class Decoding and Its Extension for Multi-Class Hypothesis Transfer Learning

Bang Zhang, Yi Wang, Yang Wang, Fang Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1075-1081

Many prevalent multi-class classification approaches can be unified and generalized by the output coding framework which usually consists of three phases: (1) coding, (2) learning binary classifiers, and (3) decoding. Most of these approaches focus on the first two phases and predefined distance function is used for decoding. In this paper, however, we propose to perform learning in coding space for more adaptive decoding, thereby improving overall performance. Ramp loss is exploited for measuring multi-class decoding error. The proposed algorithm has uniform stability. It is insensitive to data noises and scalable with large scale datasets. Generalization error bound and numerical results are given with promising outcomes.

****

Finding the Subspace Mean or Median to Fit Your Need

Tim Marrinan, J. Ross Beveridge, Bruce Draper, Michael Kirby, Chris Peterson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1082-1089

Many computer vision algorithms employ subspace models to represent data. Many o f these approaches benefit from the ability to create an average or prototype for a set of subspaces. The most popular method in these situations is the Karcher mean, also known as the Riemannian center of mass. The prevalence of the Karcher mean may lead some to assume that it provides the best average in all scenarios. However, other subspace averages that appear less frequently in the literature may be more appropriate for certain tasks. The extrinsic manifold mean, the L2-median, and the flag mean are alternative averages that can be substituted directly for the Karcher mean in many applications. This paper evaluates the characteristics and performance of these four averages on synthetic and real-world data. While the Karcher mean generalizes the Euclidean mean to the Grassman manifold, we show that the extrinsic manifold mean, the L2-median, and the flag mean behave more like medians and are therefore more robust to the presence of outliers among the subspaces being averaged. We also show that while the Karcher mean and L2-median are computed using iterative algorithms, the extrinsic manifold mean and flag mean can be found analytically and are thus orders of magnitude faster in practice. Finally, we show that the flag mean is a generalization of the extrinsic manifold mean that permits subspaces with different numbers of dimensions to be averaged. The result is a "cookbook" that maps algorithm constraints and data properties to the most appropriate subspace mean for a given application.

****

Adaptive Color Attributes for Real-Time Visual Tracking

Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, Joost van de Weijer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1090-1097

Visual tracking is a challenging problem in computer vision. Most state-of-the-art visual trackers either rely on luminance information or use simple color representations for image description. Contrary to visual tracking, for object recognition and detection, sophisticated color features when combined with luminance have shown to provide excellent performance. Due to the complexity of the tracking problem, the desired color feature should be computationally efficient, and possess a certain amount of photometric invariance while maintaining high discriminative power. This paper investigates the contribution of color in a tracking-by-detection framework. Our results suggest that color attributes provides superior performance for visual tracking. We further propose an adaptive low-dimensional variant of color attributes. Both quantitative and attribute-based evaluations are performed on 41 challenging benchmark color sequences. The proposed approach improves the baseline intensity-based tracker by 24 % in median distance precision. Furthermore, we show that our approach outperforms state-of-the-art tracking methods while running at more than 100 frames per second.

****

Local Layering for Joint Motion Estimation and Occlusion Detection

Deqing Sun, Ce Liu, Hanspeter Pfister; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1098-1105

Most motion estimation algorithms (optical flow, layered models) cannot handle large amount of occlusion in textureless regions, as motion is often initialized with no occlusion assumption despite that occlusion may be included in the final objective. To handle such situations, we propose a local layering model where motion and occlusion relationships are inferred jointly. In particular, the uncertainties of occlusion relationships are retained so that motion is inferred by considering all the possibilities of local occlusion relationships. In addition, the local layering model handles articulated objects with self-occlusion. We demonstrate that the local layering model can handle motion and occlusion well for both challenging synthetic and real sequences.

****

Realtime and Robust Hand Tracking from Depth

Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1106-1113

We present a realtime hand tracking system using a depth sensor. It tracks a fully articulated hand under large viewpoints in realtime (25 FPS on a desktop with out using a GPU) and with high accuracy (error below 10 mm). To our knowledge, it is the first system that achieves such robustness, accuracy, and speed simultaneously, as verified on challenging real data. Our system is made of several novel techniques. We model a hand simply using a number of spheres and define a fast cost function. Those are critical for realtime performance. We propose a hybrid method that combines gradient based and stochastic optimization methods to achieve fast convergence and good accuracy. We present new finger detection and hand initialization methods that greatly enhance the robustness of tracking.

****

Multi-Output Learning for Camera Relocalization

Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, Shahram Izadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1114-1121

We address the problem of estimating the pose of a cam- era relative to a known 3D scene from a single RGB-D frame. We formulate this problem as inversion of the generative rendering procedure, i.e., we want to find the camera pose corresponding to a rendering of the 3D scene model that is most similar with the observed input. This is a non-convex optimization problem with many local optima. We propose a hybrid discriminative-generative learning architecture that consists of: (i) a set of M predictors which generate M camera pose hypotheses; and (ii) a 'selector' or 'aggregator' that infers the best pose from the multiple pose hypotheses based on a similarity function. We are interested in predictors that not only produce good hypotheses but also hypotheses that are different from each other. Thus, we propose and study methods for learning 'marginally relevant' predictors, and compare their performance when used with different selection procedures. We evaluate our method on a recently released 3D reconstruction dataset with challenging camera poses, and scene variability. Experiments show that our method learns to make multiple predictions that are marginally relevant and can effectively select an accurate prediction. Furthermore, our method outperforms the state-of-the-art discriminative approach for camera relocalization.

****

MAP Visibility Estimation for Large-Scale Dynamic 3D Reconstruction

Hanbyul Joo, Hyun Soo Park, Yaser Sheikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1122-1129

Many traditional challenges in reconstructing 3D motion, such as matching across wide baselines and handling occlusion, reduce in significance as the number of unique viewpoints increases. However, to obtain this benefit, a new challenge arises: estimating precisely which cameras observe which points at each instant in time. We present a maximum a posteriori (MAP) estimate of the time-varying visibility of the target points to reconstruct the 3D motion of an event from a large number of cameras. Our algorithm takes, as input, camera poses and image sequences, and outputs the time-varying set of the cameras in which a target patch is visibile and its reconstructed trajectory. We model visibility estimation as a MAP estimate by incorporating various cues including photometric consistency, motion consistency, and geometric consistency, in conjunction with a prior that rewards consistent visibilities in proximal cameras. An optimal estimate of visibility is obtained by finding the minimum cut of a capacitated graph over cameras. We demonstrate that our method estimates visibility with greater accuracy, and increases tracking performance producing longer trajectories, at more locations, and at higher accuracies than methods that ignore visibility or use photometric consistency alone.

****

Multi-Object Tracking via Constrained Sequential Labeling

Sheng Chen, Alan Fern, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1130-1137

This paper presents a new approach to tracking people in crowded scenes, where people are subject to long-term (partial) occlusions and may assume varying postures and articulations. In such videos, detection-based trackers give poor performance since detecting people occurrences is not reliable, and common assumptions about locally smooth trajectories do not hold. Rather, we use temporal mid-leve l features (e.g., supervoxels or dense point trajectories) as a more coherent spatiotemporal basis for handling occlusion and pose variations.Thus, we formulate tracking as labeling mid-level features by object identifiers, and specify a new approach, called constrained sequential labeling (CSL), for performing this labeling. CSL uses a cost function to sequentially assign labels while respecting the implications of hard constraints computed via constraint propagation. A key feature of this approach is that it allows for the use of flexible cost functions and constraints that capture complex dependencies that cannot be represented in standard network-flow formulations. To exploit this flexibility we describe how to learn constraints and give a provably correct learning algorithms for cost functions that achieves finitetime convergence at a rate that improves with the strength of the constraints. Our experimental results indicate that CSL outperforms the state-of-the-art on challenging real-world videos of volleyball, basketball, and pedestrians walking.

****

A Primal-Dual Algorithm for Higher-Order Multilabel Markov Random Fields

Alexander Fix, Chen Wang, Ramin Zabih; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1138-1145

Graph cuts method such as a-expansion [4] and fusion moves [22] have been successful at solving many optimization problems in computer vision. Higher-order Markov Random Fields (MRF's), which are important for numerous applications, have proven to be very difficult, especially for multilabel MRF's (i.e. more than 2 labels). In this paper we propose a new primal-dual energy minimization method for arbitrary higher-order multilabel MRF's. Primal-dual methods provide guaranteed approximation bounds, and can exploit information in the dual variables to improve their efficiency. Our algorithm generalizes the PD3 [19] technique for first-order MRFs, and relies on a variant of max-flow that can exactly optimize certain higher-order binary MRF's [14]. We provide approximation bounds similar to PD3 [19], and the method is fast in practice. It can optimize non-submodular MRF's, and additionally can in- corporate problem-specific knowledge in the form of fusion proposals. We compare experimentally against the existing approaches that can efficiently handle these difficult energy functions [6, 10, 11]. For higher-order denoising and stereo MRF's, we produce lower energy while running significantly faster.

****

Energy Based Multi-model Fitting & Matching for 3D Reconstruction

Hossam Isack, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1146-1153

Standard geometric model fitting methods take as an input a fixed set of feature pairs greedily matched based only on their appearances. Inadvertently, many valid matches are discarded due to repetitive texture or large baseline between view points. To address this problem, matching should consider both feature appearances and geometric fitting errors. We jointly solve feature matching and multi-model fitting problems by optimizing one energy. The formulation is based on our generalization of the assignment problem and its efficient min-cost-max-flow solver. Our approach significantly increases the number of correctly matched features, improves the accuracy of fitted models, and is robust to larger baselines.

****

Submodularization for Binary Pairwise Energies

Lena Gorelick, Yuri Boykov, Olga Veksler, Ismail Ben Ayed, Andrew Delong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1154-1161

Many computer vision problems require optimization of binary non-submodular energies. We propose a general optimization framework based on local submodular approximations (LSA). Unlike standard LP relaxation methods that linearize the whole energy globally, our approach iteratively approximates the energies locally. On the other hand, unlike standard local optimization methods (e.g. gradient descent or projection techniques) we use non-linear submodular approximations and optimize them without leaving the domain of integer solutions. We discuss two specific LSA algorithms based on trust region and auxiliary function principles, LSA

-TR and LSA-AUX. These methods obtain state-of-the-art results on a wide range of applications outperforming many standard techniques such as LBP, QPBO, and TRWS. While our paper is focused on pairwise energies, our ideas extend to higher-order problems. The code is available online

****

Maximum Persistency in Energy Minimization

Alexander Shekhovtsov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1162-1169

We consider discrete pairwise energy minimization problem (weighted constraint satisfaction, max-sum labeling) and methods that identify a globally optimal partial assignment of variables. When finding a complete optimal assignment is intractable, determining optimal values for a part of variables is an interesting possibility. Existing methods are based on different sufficient conditions. We propose a new sufficient condition for partial optimality which is: (1) verifiable in polynomial time (2) invariant to reparametrization of the problem and permutation of labels and (3) includes many existing sufficient conditions as special cases. We pose the problem of finding the maximum optimal partial assignment identifiable by the new sufficient condition. A polynomial method is proposed which is guaranteed to assign same or larger part of variables. The core of the method is a specially constructed linear program that identifies persistent assignments in an arbitrary multi-label setting.

****

Partial Optimality by Pruning for MAP-inference with General Graphical Models

Paul Swoboda, Bogdan Savchynskyy, Jorg H. Kappes, Christoph Schnorr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1170-1177

We consider the energy minimization problem for undirected graphical models, also known as MAP-inference problem for Markov random fields which is NP-hard in general. We propose a novel polynomial time algorithm to obtain a part of its optimal nonrelaxed integral solution. Our algorithm is initialized with variables taking integral values in the solution of a convex relaxation of the MAP-inference problem and iteratively prunes those, which do not satisfy our criterion for partial optimality. We show that our pruning strategy is in a certain sense theoretically optimal. Also empirically our method outperforms previous approaches in terms of the number of persistently labelled variables. The method is very general, as it is applicable to models with arbitrary factors of an arbitrary order and can employ any solver for the considered relaxed problem. Our method's runtime is determined by the runtime of the convex relaxation solver for the MAP-inference problem.

****

Scene Labeling Using Beam Search Under Mutex Constraints

Anirban Roy, Sinisa Todorovic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1178-1185

This paper addresses the problem of assigning object class labels to image pixels. Following recent holistic formulations, we cast scene labeling as inference of a conditional random field (CRF) grounded onto superpixels. The CRF inference is specified as quadratic program (QP) with mutual exclusion (mutex) constraints on class label assignments. The QP is solved using a beam search (BS), which is well-suited for scene labeling, because it explicitly accounts for spatial extents of objects; conforms to inconsistency constraints from domain knowledge; and has low computational costs. BS gradually builds a search tree whose nodes correspond to candidate scene labelings. Successor nodes are repeatedly generated from a select set of their parent nodes until convergence. We prove that our BS efficiently maximizes the QP objective of CRF inference. Effectiveness of our BS for scene labeling is evaluated on the benchmark MSRC, Stanford Backgroud, PASCAL VOC 2009 and 2010 datasets.

****

Persistent Tracking for Wide Area Aerial Surveillance

Jan Prokaj, Gerard Medioni; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1186-1193

Persistent surveillance of large geographic areas from unmanned aerial vehicles allows us to learn much about the daily activities in the region of interest. Nearly all of the approaches addressing tracking in this imagery are detection-based and rely on background subtraction or frame differencing to provide detections. This, however, makes it difficult to track targets once they slow down or stop, which is not acceptable for persistent tracking, our goal. We present a multiple target tracking approach that does not exclusively rely on background subtraction and is better able to track targets through stops. It accomplishes this by effectively running two trackers in parallel: one based on detections from background subtraction providing target initialization and reacquisition, and one based on a target state regressor providing frame to frame tracking. We evaluated the proposed approach on a long sequence from a wide area aerial imagery dataset, and the results show improved object detection rates and ID-switch rates with limited increases in false alarms compared to the competition.

****

Multi-Cue Visual Tracking Using Robust Feature-Level Fusion Based on Joint Sparse Representation

Xiangyuan Lan, Andy J. Ma, Pong C. Yuen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1194-1201

The use of multiple features for tracking has been proved as an effective approach because limitation of each feature could be compensated. Since different types of variations such as illumination, occlusion and pose may happen in a video sequence, especially long sequence videos, how to dynamically select the appropriate features is one of the key problems in this approach. To address this issue in multicue visual tracking, this paper proposes a new joint sparse representation model for robust feature-level fusion. The proposed method dynamically removes unreliable features to be fused for tracking by using the advantages of sparse representation. As a result, robust tracking performance is obtained. Experimental results on publicly available videos show that the proposed method outperforms both existing sparse representation based and fusion-based trackers.

****

Multi-Forest Tracker: A Chameleon in Tracking

David J. Tan, Slobodan Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1202-1209

In this paper, we address the problem of object tracking in intensity images and depth data. We propose a generic framework that can be used either for tracking 2D templates in intensity images or for tracking 3D objects in depth images. To overcome problems like partial occlusions, strong illumination changes and motion blur, that notoriously make energy minimization-based tracking methods get trapped in a local minimum, we propose a learning-based method that is robust to all these problems. We use random forests to learn the relation between the parameters that defines the object's motion, and the changes they induce on the image intensities or the point cloud of the template. It follows that, to track the template when it moves, we use the changes on the image intensities or point cloud to predict the parameters of this motion. Our algorithm has an extremely fast tracking performance running at less than 2 ms per frame, and is robust to partial occlusions. Moreover, it demonstrates robustness to strong illumination changes when tracking templates using intensity images, and robustness in tracking 3D objects from arbitrary viewpoints even in the presence of motion blur that causes missing or erroneous data in depth images. Extensive experimental evaluation and comparison to the related approaches strongly demonstrates the benefits of our method.

****

Rigid Motion Segmentation using Randomized Voting

Heechul Jung, Jeongwoo Ju, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1210-1217

In this paper, we propose a novel rigid motion segmentation algorithm called randomized voting (RV). This algorithm is based on epipolar geometry, and computes a score using the distance between the feature point and the corresponding epipolar line. This score is accumulated and utilized for final grouping. Our algorit hm basically deals with two frames, so it is also applicable to the two-view motion segmentation problem. For evaluation of our algorithm, Hopkins 155 dataset, which is a representative test set for rigid motion segmentation, is adopted; it consists of two and three rigid motions. Our algorithm has provided the most accurate motion segmentation results among all of the state-of-the-art algorithms. The average error rate is 0.77%. In addition, when there is measurement noise, our algorithm is comparable with other state-of-the-art algorithms.

****

Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning
Seung-Hwan Bae, Kuk-Jin Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1218-1225

Online multi-object tracking aims at producing complete tracks of multiple objects using the information accumulated up to the present moment. It still remains a difficult problem in complex scenes, because of frequent occlusion by clutter or other objects, similar appearances of different objects, and other factors. In this paper, we propose a robust online multi-object tracking method that can handle these difficulties effectively. We first propose the tracklet confidence using the detectability and continuity of a tracklet, and formulate a multi-object tracking problem based on the tracklet confidence. The multi-object tracking problem is then solved by associating tracklets in different ways according to their confidence values. Based on this strategy, tracklets sequentially grow with online-provided detections, and fragmented tracklets are linked up with others without any iterative and expensive associations. Here, for reliable association between tracklets and detections, we also propose a novel online learning method using an incremental linear discriminant analysis for discriminating the appearances of objects. By exploiting the proposed learning method, tracklet association can be successfully achieved even under severe occlusion. Experiments with challenging public datasets show distinct performance improvement over other batch and online tracking methods.

****

Pyramid-based Visual Tracking Using Sparsity Represented Mean Transform
Zhe Zhang, Kin Hong Wong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1226-1233

In this paper, we propose a robust method for visual tracking relying on mean shift, sparse coding and spatial pyramids. Firstly, we extend the original mean shift approach to handle orientation space and scale space and name this new method as mean transform. The mean transform method estimates the motion, including the location, orientation and scale, of the interested object window simultaneously and effectively. Secondly, a pixel-wise dense patch sampling technique and a region-wise trivial template designing scheme are introduced which enable our approach to run very accurately and efficiently. In addition, instead of using either holistic representation or local representation only, we apply spatial pyramids by combining these two representations into our approach to deal with partial occlusion problems robustly. Observed from the experimental results, our approach outperforms state-of-the-art methods in many benchmark sequences.

****

Tracklet Association with Online Target-Specific Metric Learning
Bing Wang, Gang Wang, Kap Luk Chan, Li Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1234-1241

This paper presents a novel introduction of online target-specific metric learning in track fragment (tracklet) association by network flow optimization for long-term multi-person tracking. Different from other network flow formulation, each node in our network represents a tracklet, and each edge represents the likelihood of   neighboring tracklets belonging to the same trajectory as measured by our proposed affinity score. In our method, target-specific similarity metrics are learned, which give rise to the appearance-based models used in the tracklet affinity estimation. Trajectory-based tracklets are refined by using the learned metrics to account for appearance consistency and to identify reliable tracklets. The metrics are then re-learned using reliable tracklets for computing track let affinity scores. Long-term trajectories are then obtained through network flow optimization. Occlusions and missed detections are handled by a trajectory completion step. Our method is effective for long-term tracking even when the targets are spatially close or completely occluded by others. We validate our proposed framework on several public datasets and show that it outperforms several state of art methods.

****

An Online Learned Elementary Grouping Model for Multi-target Tracking

Xiaojing Chen, Zhen Qin, Le An, Bir Bhanu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1242-1249

We introduce an online approach to learn possible elementary groups (groups that contain only two targets) for inferring high level context that can be used to improve multi-target tracking in a data-association based framework. Unlike most existing association-based tracking approaches that use only low level information (e.g., time, appearance, and motion) to build the affinity model and consider each target as an independent agent, we online learn social grouping behavior to provide additional information for producing more robust tracklets affinities. Social grouping behavior of pairwise targets is first learned from confident tracklets and encoded in a disjoint grouping graph. The grouping graph is further completed with the help of group tracking. The proposed method is efficient, handles group merge and split, and can be easily integrated into any basic affinity model. We evaluate our approach on two public datasets, and show significant improvements compared with state-of-the-art methods.

****

Diversity-Enhanced Condensation Algorithm and Its Application for Robust and Accurate Endoscope Three-Dimensional Motion Tracking

Xiongbiao Luo, Ying Wan, Xiangjian He, Jie Yang, Kensaku Mori; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1250-1257

The paper proposes a diversity-enhanced condensation algorithm to address the particle impoverishment problem which stochastic filtering usually suffers from. The particle diversity plays an important role as it affects the performance of filtering. Although the condensation algorithm is widely used in computer vision, it easily gets trapped in local minima due to the particle degeneracy. We introduce a modified evolutionary computing method, adaptive differential evolution, to resolve the particle impoverishment under a proper size of particle population. We apply our proposed method to endoscope tracking for estimating three-dimensional motion of the endoscopic camera. The experimental results demonstrate that our proposed method offers more robust and accurate tracking than previous methods. The current tracking smoothness and error were significantly reduced from (3.7, 4.8) to (2.3 mm, 3.2 mm), which approximates the clinical requirement of 3.0 mm.

****

Partial Occlusion Handling for Visual Tracking via Robust Part Matching

Tianzhu Zhang, Kui Jia, Changsheng Xu, Yi Ma, Narendra Ahuja; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1258-1265

Part-based visual tracking is advantageous due to its robustness against partial occlusion. However, how to effectively exploit the confidence scores of individual parts to construct a robust tracker is still a challenging problem. In this paper, we address this problem by simultaneously matching parts in each of multiple frames, which is realized by a locality-constrained low-rank sparse learning method that establishes multi-frame part correspondences through optimization of partial permutation matrices. The proposed part matching tracker (PMT) has a number of attractive properties. (1) It exploits the spatial-temporal localityconstrained property for robust part matching. (2) It matches local parts from multiple frames jointly by considering their low-rank and sparse structure information, which can effectively handle part appearance variations due to occlusion or noise. (3) The proposed PMT model has the inbuilt mechanism of leveraging multi-mode target templates, so that the dilemma of template updating when encounterin g occlusion in tracking can be better handled. This contrasts with existing methods that only do part matching between a pair of frames. We evaluate PMT and compare with 10 popular state-of-the-art methods on challenging benchmarks. Experimental results show that PMT consistently outperform these existing trackers.

****

Speeding Up Tracking by Ignoring Features

Lu Zhang, Hamdi Dibeklioglu, Laurens van der Maaten; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1266-1273

Most modern object trackers combine a motion prior with sliding-window detection, using binary classifiers that predict the presence of the target object based on histogram features. Although the accuracy of such trackers is generally very good, they are often impractical because of their high computational requirements. To resolve this problem, the paper presents a new approach that limits the computational costs of trackers by ignoring features in image regions that --- after inspecting a few features --- are unlikely to contain the target object. To this end, we derive an upper bound on the probability that a location is most likely to contain the target object, and we ignore (features in) locations for which this upper bound is small. We demonstrate the effectiveness of our new approach in experiments with model-free and model-based trackers that use linear models in combination with HOG features. The results of our experiments demonstrate that our approach allows us to reduce the average number of inspected features by up to 90% without affecting the accuracy of the tracker.

****

Subspace Tracking under Dynamic Dimensionality for Online Background Subtraction

Matthew Berger, Lee M. Seversky; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1274-1281

Long-term modeling of background motion in videos is an important and challenging problem used in numerous applications such as segmentation and event recognition. A major challenge in modeling the background from point trajectories lies in dealing with the variable length duration of trajectories, which can be due to such factors as trajectories entering and leaving the frame or occlusion from different depth layers. This work proposes an online method for background modeling of dynamic point trajectories via tracking of a linear subspace describing the background motion. To cope with variability in trajectory durations, we cast subspace tracking as an instance of subspace estimation under missing data, using a least-absolute deviations formulation to robustly estimate the background in the presence of arbitrary foreground motion. Relative to previous works, our approach is very fast and scales to arbitrarily long videos as our method processes new frames sequentially as they arrive.

****

Multiple Target Tracking Based on Undirected Hierarchical Relation Hypergraph

Longyin Wen, Wenbo Li, Junjie Yan, Zhen Lei, Dong Yi, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1282-1289

Multi-target tracking is an interesting but challenging task in computer vision field. Most previous data association based methods merely consider the relationships (e.g. appearance and motion pattern similarities) between detections in local limited temporal domain, leading to their difficulties in handling long-term occlusion and distinguishing the spatially close targets with similar appearance in crowded scenes. In this paper, a novel data association approach based on undirected hierarchical relation hypergraph is proposed, which formulates the tracking task as a hierarchical dense neighborhoods searching problem on the dynamically constructed undirected affinity graph. The relationships between different detections across the spatiotemporal domain are considered in a high-order way, which makes the tracker robust to the spatially close targets with similar appearance. Meanwhile, the hierarchical design of the optimization process fuels our tracker to long-term occlusion with more robustness. Extensive experiments on various challenging datasets (i.e. PETS2009 dataset, ParkingLot), including both low and high density sequences, demonstrate that the proposed method performs favorably against the state-of-the-art methods.

****

Bi-label Propagation for Generic Multiple Object Tracking

Wenhan Luo, Tae-Kyun Kim, Bjorn Stenger, Xiaowei Zhao, Roberto Cipolla; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1290-1297

In this paper, we propose a label propagation framework to handle the multiple object tracking (MOT) problem for a generic object type (cf. pedestrian tracking). Given a target object by an initial bounding box, all objects of the same type are localized together with their identities. We treat this as a problem of propagating bi-labels, i.e. a binary class label for detection and individual object labels for tracking. To propagate the class label, we adopt clustered Multiple Task Learning (cMTL) while enforcing spatio-temporal consistency and show that this improves the performance when given limited training data. To track objects, we propagate labels from trajectories to detections based on affinity using appearance, motion, and context. Experiments on public and challenging new sequences show that the proposed method improves over the current state of the art on this task.

****

A Probabilistic Framework for Multitarget Tracking with Mutual Occlusions

Menglong Yang, Yiguang Liu, Longyin Wen, Zhisheng You, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1298-1305

Mutual occlusions among targets can cause track loss or target position deviation, because the observation likelihood of an occluded target may vanish even when we have the estimated location of the target. This paper presents a novel probability framework for multitarget tracking with mutual occlusions. The primary contribution of this work is the introduction of a vectorial occlusion variable as part of the solution. The occlusion variable describes occlusion states of the targets. This forms the basis of the proposed probability framework, with the following further contributions: 1) Likelihood: A new observation likelihood model is presented, in which the likelihood of an occluded target is computed by referring to both of the occluded and oc-cluding targets. 2) Priori: Markov random field (MRF) is used to model the occlusion priori such that less likely "circular" or "cascading" types of occlusions have lower priori probabilities. Both the occlusion priori and the motion priori take into consideration the state of occlusion. 3) Optimization: A realtime RJMCMC-based algorithm with a newmove type called "occlusion state update" is presented. Experimental results show that the proposed framework can handle occlusions well, even including long-duration full occlusions, which may cause tracking failures in the traditional methods.

****

Occlusion Geodesics for Online Multi-Object Tracking

Horst Possegger, Thomas Mauthner, Peter M. Roth, Horst Bischof; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1306-1313

Robust multi-object tracking-by-detection requires the correct assignment of noisy detection results to object trajectories. We address this problem by proposing an online approach based on the observation that object detectors primarily fail if objects are significantly occluded. In contrast to most existing work, we only rely on geometric information to efficiently overcome detection failures.
In particular, we exploit the spatio-temporal evolution of occlusion regions, detector reliability, and target motion prediction to robustly handle missed detections. In combination with a conservative association scheme for visible objects, this allows for real-time tracking of multiple objects from a single static camera, even in complex scenarios. Our evaluations on publicly available multi-object tracking benchmark datasets demonstrate favorable performance compared to the state-of-the-art in online and offline multi-object tracking.

****

Efficient Nonlinear Markov Models for Human Motion

Andreas M. Lehrmann, Peter V. Gehler, Sebastian Nowozin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1314-13

Dynamic Bayesian networks such as Hidden Markov Models (HMMs) are successfully used as probabilistic models for human motion. The use of hidden variables makes them expressive models, but inference is only approximate and requires procedures such as particle filters or Markov chain Monte Carlo methods. In this work we propose to instead use simple Markov models that only model observed quantities. We retain a highly expressive dynamic model by using interactions that are nonlinear and non-parametric. A presentation of our approach in terms of latent variables shows logarithmic growth for the computation of exact log-likelihoods in the number of latent states. We validate our model on human motion capture data and demonstrate state-of-the-art performance on action recognition and motion completion tasks.

****

A Compositional Model for Low-Dimensional Image Set Representation

Hossein Mobahi, Ce Liu, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1322-1329

Learning a low-dimensional representation of images is useful for various applications in graphics and computer vision. Existing solutions either require manually specified landmarks for corresponding points in the images, or are restricted to specific objects or shape deformations. This paper alleviates these limitations by imposing a specific model for generating images; the nested composition of color, shape, and appearance. We show that each component can be approximated by a low-dimensional subspace when the others are factored out. Our formulation allows for efficient learning and experiments show encouraging results.

****

A Principled Approach for Coarse-to-Fine MAP Inference

Christopher Zach; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1330-1337

In this work we reconsider labeling problems with (virtually) continuous state spaces, which are of relevance in low level computer vision. In order to cope with such huge state spaces multi-scale methods have been proposed to approximately solve such labeling tasks. Although performing well in many cases, these methods do usually not come with any guarantees on the returned solution. A general and principled approach to solve labeling problems is based on the well-known linear programming relaxation, which appears to be prohibitive for large state spaces at the first glance. We demonstrate that a coarse-to-fine exploration strategy in the label space is able to optimize the LP relaxation for non-trivial problem instances with reasonable run-times and moderate memory requirements.

****

Fast Approximate Inference in Higher Order MRF-MAP Labeling Problems

Chetan Arora, Subhashis Banerjee, Prem Kalra, S.N. Maheshwari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1338-1345

Use of higher order clique potentials for modeling inference problems has exploded in last few years. The algorithmic schemes proposed so far do not scale well with increasing clique size, thus limiting their use to cliques of size at most 4 in practice. Generic Cuts (GC) of Arora et al. [9] shows that when potentials are submodular, inference problems can be solved optimally in polynomial time for fixed size cliques. In this paper we report an algorithm called Approximate Cuts (AC) which uses a generalization of the gadget of GC and provides an approximate solution to inference in 2-label MRF-MAP problems with cliques of size $k \geq 2$. The algorithm gives optimal solution for submodular potentials. When potentials are non-submodular, we show that important properties such as weak persistency hold for solution inferred by AC. AC is a polynomial time primal dual approximation algorithm for fixed clique size. We show experimentally that AC not only provides significantly better solutions in practice, it is an order of magnitude faster than message passing schemes like Dual Decomposition [19] and GTRWS [17] or Reduction based techniques like [10, 13, 14].

****

Multi Label Generic Cuts: Optimal Inference in Multi Label Multi Clique MRF-MAP

Problems

Chetan Arora, S.N. Maheshwari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1346-1353

We propose an algorithm called Multi Label Generic Cuts (MLGC) for computing optimal solutions to MRF-MAP problems with submodular multi label multi-clique potentials. A transformation is introduced to convert a m-label k-clique problem to an equivalent 2-label (mk)-clique problem. We show that if the original multi-label problem is submodular then the transformed 2-label multi-clique problem is also submodular. We exploit sparseness in the feasible configurations of the transformed 2-label problem to suggest an improvement to Generic Cuts [3] to solve the 2-label problems efficiently. The algorithm runs in time O(m^k n^3) in the worst case (n is the number of pixels) generalizing O(2^k n^3) running time of Generic Cuts. We show experimentally that MLGC is an order of magnitude faster than the current state of the art [17, 20]. While the result of MLGC is optimal for submodular clique potential it is significantly better than the compared methods even for problems with non-submodular clique potential.

****

Scanline Sampler without Detailed Balance: An Efficient MCMC for MRF Optimization

Wonsik Kim, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1354-1361

Markov chain Monte Carlo (MCMC) is an elegant tool, widely used in variety of areas. In computer vision, it has been used for the inference on the Markov random field model (MRF). However, MCMC less concerned than other deterministic approaches although it converges to global optimal solution in theory. The major obstacle is its slow convergence. To come up with faster sampling method, we investigate two ideas: breaking detailed balance and updating multiple nodes at a time. Although detailed balance is considered to be essential element of MCMC, it actually is not the necessary condition for the convergence. In addition, exploiting the structure of MRF, we introduce a new kernel which updates multiple nodes in a scanline rather than a single node. Those two ideas are integrated in a novel way to develop an efficient method called scanline sampler without detailed balance. In experimental section, we apply our method to the OpenGM2 benchmark of MRF optimization and show the proposed method achieves faster convergence than the conventional approaches.

****

Higher-Order Clique Reduction Without Auxiliary Variables

Hiroshi Ishikawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1362-1369

We introduce a method to reduce most higher-order terms of Markov Random Fields with binary labels into lower-order ones without introducing any new variables, while keeping the minimizer of the energy unchanged. While the method does not reduce all terms, it can be used with existing techniques that transformsarbitrary terms (by introducing auxiliary variables) and improve the speed. The method eliminates a higher-order term in the polynomial representation of the energy by finding the value assignment to the variables involved that cannot be part of a global minimizer and increasing the potential value only when that particular combination occurs by the exact amount that makes the potential of lower order. We also introduce a faster approximation that forego the guarantee of exact equivalence of minimizer in favor of speed. With experiments on the same field of experts dataset used in previous work, we show that the roof-dual algorithm after the reduction labels significantly more variables and the energy converges more rapidly.

****

Topic Modeling of Multimodal Data: An Autoregressive Approach

Yin Zheng, Yu-Jin Zhang, Hugo Larochelle; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1370-1377

Topic modeling based on latent Dirichlet allocation (LDA) has been a framework of choice to deal with multimodal data, such as in image annotation tasks. Recently, a new type of topic model called the Document Neural Autoregressive Distribu tion Estimator (DocNADE) was proposed and demonstrated state-of-the-art performance for text document modeling. In this work, we show how to successfully apply and extend this model to multimodal data, such as simultaneous image classification and annotation. Specifically, we propose SupDocNADE, a supervised extension of DocNADE, that increases the discriminative power of the hidden topic features by incorporating label information into the training objective of the model and show how to employ SupDocNADE to learn a joint representation from image visual words, annotation words and class label information. We also describe how to leverage information about the spatial position of the visual words for SupDocNADE to achieve better performance in a simple, yet effective manner. We test our model on the LabelMe and UIUC-Sports datasets and show that it compares favorably to other topic models such as the supervised variant of LDA and a Spatial Matching Pyramid (SPM) approach.

****

Model Transport: Towards Scalable Transfer Learning on Manifolds

Oren Freifeld, Soren Hauberg, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1378-1385

We consider the intersection of two research fields: transfer learning and statistics on manifolds. In particular, we consider, for manifold-valued data, transfer learning of tangent-space models such as Gaussians distributions, PCA, regression, or classifiers. Though one would hope to simply use ordinary Rn -transfer learning ideas, the manifold structure prevents it. We overcome this by basing our method on inner-product-preserving parallel transport, a well-known tool widely used in other problems of statistics on manifolds in computer vision. At first, this straight-forward idea seems to suffer from an obvious shortcoming: Transporting large datasets is prohibitively expensive, hindering scalability. Fortunately, with our approach, we never transport data. Rather, we show how the statistical models themselves can be transported, and prove that for the tangent-space models above, the transport "commutes" with learning. Consequently, our compact framework, applicable to a large class of manifolds, is not restricted by the size of either the training or test sets. We demonstrate the approach by transferring PCA and logistic-regression models of real-world data involving 3D shapes and image descriptors.

****

Learning Fine-grained Image Similarity with Deep Ranking

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, Ying Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1386-1393

Learning fine-grained image similarity is a challenging task. It needs to capture between-class and within-class image differences. This paper proposes a deep ranking model that employs deep learning techniques to learn similarity metric directly from images. It has higher learning capability than models based on hand-crafted features. A novel multiscale network structure has been developed to describe the images effectively. An efficient triplet sampling algorithm is also proposed to learn the model with distributed asynchronized stochastic gradient. Extensive experiments show that the proposed algorithm outperforms models based on hand-crafted visual features and deep classification models.

****

Attributed Graph Mining and Matching: An Attempt to Define and Extract Soft Attributed Patterns

Quanshi Zhang, Xuan Song, Xiaowei Shao, Huijing Zhao, Ryosuke Shibasaki; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1394-1401

Graph matching and graph mining are two typical areas in artificial intelligence. In this paper, we define the soft attributed pattern (SAP) to describe the common subgraph pattern among a set of attributed relational graphs (ARGs), considering both the graphical structure and graph attributes. We propose a direct solution to extract the SAP with the maximal graph size without node enumeration. Given an initial graph template and a number of ARGs, we modify the graph template into the maximal SAP among the ARGs in an unsupervised fashion. The maximal SAP extraction is equivalent to learning a graphical model (i.e. an object model) from large ARGs (i.e. cluttered RGB/RGB-D images) for graph matching, which extends the concept of "unsupervised learning for graph matching." Furthermore, this study can be also regarded as the first known approach to formulating "maximal graph mining" in the graph domain of ARGs. Our method exhibits superior performance on RGB and RGB-D images.

****

Deep Fisher Kernels - End to End Learning of the Fisher Kernel GMM Parameters

Vladyslav Sydorov, Mayu Sakurada, Christoph H. Lampert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1402-1409

Fisher Kernels and Deep Learning were two developments with significant impact on large-scale object categorization in the last years. Both approaches were shown to achieve state-of-the-art results on large-scale object categorization datasets, such as ImageNet. Conceptually, however, they are perceived as very different and it is not uncommon for heated debates to spring up when advocates of both paradigms meet at conferences or workshops. In this work, we emphasize the similarities between both architectures rather than their differences and we argue that such a unified view allows us to transfer ideas from one domain to the other. As a concrete example we introduce a method for learning a support vector machine classifier with Fisher kernel at the same time as a task-specific data representation. We reinterpret the setting as a multi-layer feed forward network. Its final layer is the classifier, parameterized by a weight vector, and the two previous layers compute Fisher vectors, parameterized by the coefficients of a Gaussian mixture model. We introduce a gradient descent based learning algorithm that, in contrast to other feature learning techniques, is not just derived from intuition or biological analogy, but has a theoretical justification in the framework of statistical learning theory. Our experiments show that the new training procedure leads to significant improvements in classification accuracy while preserving the modularity and geometric interpretability of a support vector machine setup.

****

Transfer Joint Matching for Unsupervised Domain Adaptation

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, Philip S. Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1410-1417

Visual domain adaptation, which learns an accurate classifier for a new domain using labeled images from an old domain, has shown promising value in computer vision yet still been a challenging problem. Most prior works have explored two learning strategies independently for domain adaptation: feature matching and instance reweighting. In this paper, we show that both strategies are important and inevitable when the domain difference is substantially large. We therefore put forward a novel Transfer Joint Matching (TJM) approach to model them in a unified optimization problem. Specifically, TJM aims to reduce the domain difference by jointly matching the features and reweighting the instances across domains in a principled dimensionality reduction procedure, and construct new feature representation that is invariant to both the distribution difference and the irrelevant instances. Comprehensive experimental results verify that TJM can significantly outperform competitive methods for cross-domain image recognition problems.

****

Recognizing RGB Images by Learning from RGB-D Data

Lin Chen, Wen Li, Dong Xu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1418-1425

In this work, we propose a new framework for recognizing RGB images captured by the conventional cameras by leveraging a set of labeled RGB-D data, in which the depth features can be additionally extracted from the depth images. We formulate this task as a new unsupervised domain adaptation (UDA) problem, in which we aim to take advantage of the additional depth features in the source domain and also cope with the data distribution mismatch between the source and target domains. To effectively utilize the additional depth features, we seek two optimal p

rojection matrices to map the samples from both domains into a common space by preserving as much as possible the correlations between the visual features and depth features. To effectively employ the training samples from the source domain for learning the target classifier, we reduce the data distribution mismatch by minimizing the Maximum Mean Discrepancy (MMD) criterion, which compares the data distributions for each type of feature in the common space. Based on the above two motivations, we propose a new SVM based objective function to simultaneously learn the two projection matrices and the optimal target classifier in order to well separate the source samples from different classes when using each type of feature in the common space. An efficient alternating optimization algorithm is developed to solve our new objective function. Comprehensive experiments for object recognition and gender recognition demonstrate the effectiveness of our proposed approach for recognizing RGB images by learning from RGB-D data.
****

Instance-weighted Transfer Learning of Active Appearance Models

Daniel Haase, Erik Rodner, Joachim Denzler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1426-1433

There has been a lot of work on face modeling, analysis, and landmark detection, with Active Appearance Models being one of the most successful techniques. A major drawback of these models is the large number of detailed annotated training examples needed for learning. Therefore, we present a transfer learning method that is able to learn from related training data using an instance-weighted transfer technique. Our method is derived using a generalization of importance sampling and in contrast to previous work we explicitly try to tackle the transfer already during learning instead of adapting the fitting process. In our studied application of face landmark detection, we efficiently transfer facial expressions from other human individuals and are thus able to learn a precise face Active Appearance Model only from neutral faces of a single individual. Our approach is evaluated on two common face datasets and outperforms previous transfer methods.
****

Scalable Multitask Representation Learning for Scene Classification

Maksim Lapin, Bernt Schiele, Matthias Hein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1434-1441

The underlying idea of multitask learning is that learning tasks jointly is better than learning each task individually. In particular, if only a few training examples are available for each task, sharing a jointly trained representation improves classification performance. In this paper, we propose a novel multitask learning method that learns a low-dimensional representation jointly with the corresponding classifiers, which are then able to profit from the latent inter-class correlations. Our method scales with respect to the original feature dimension and can be used with high-dimensional image descriptors such as the Fisher Vector. Furthermore, it consistently outperforms the current state of the art on the SUN397 scene classification benchmark with varying amounts of training data.
****

Learning to Learn, from Transfer Learning to Domain Adaptation: A Unifying Perspective

Novi Patricia, Barbara Caputo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1442-1449

The transfer learning and domain adaptation problems originate from a distribution mismatch between the source and target data distribution. The causes of such mismatch are traditionally considered different. Thus, transfer learning and domain adaptation algorithms are designed to address different issues, and cannot be used in both settings unless substantially modified. Still, one might argue that these problems are just different declinations of learning to learn, i.e. the ability to leverage over prior knowledge when attempting to solve a new task. We propose a learning to learn framework able to leverage over source data regardless of the origin of the distribution mismatch. We consider prior models as experts, and use their output confidence value as features. We use them to build the new target model, combined with the features from the target data through a high-level cue integration scheme. This results in a class of algorithms usable in

a plug-and-play fashion over any learning to learn scenario, from binary and multi-class transfer learning to single and multiple source domain adaptation settings. Experiments on several public datasets show that our approach consistently achieves the state of the art.
****

Constructing Robust Affinity Graphs for Spectral Clustering
Xiatian Zhu, Chen Change Loy, Shaogang Gong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1450-1457
Spectral clustering requires robust and meaningful affinity graphs as input in order to form clusters with desired structures that can well support human intuition. To construct such affinity graphs is non-trivial due to the ambiguity and uncertainty inherent in the raw data. In contrast to most existing clustering methods that typically employ all available features to construct affinity matrices with the Euclidean distance, which is often not an accurate representation of the underlying data structures, we propose a novel unsupervised approach to generating more robust affinity graphs via identifying and exploiting discriminative features for improving spectral clustering. Specifically, our model is capable of capturing and combining subtle similarity information distributed over discriminative feature subspaces for more accurately revealing the latent data distribution and thereby leading to improved data clustering, especially with heterogeneous data sources. We demonstrate the efficacy of the proposed approach on challenging image and video datasets.
****

A Fast and Robust Algorithm to Count Topologically Persistent Holes in Noisy Clouds
Vitaliy Kurlin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1458-1463
Preprocessing a 2D image often produces a noisy cloud of interest points. We study the problem of counting holes in noisy clouds in the plane. The holes in a given cloud are quantified by the topological persistence of their boundary contours when the cloud is analyzed at all possible scales. We design the algorithm to count holes that are most persistent in the filtration of offsets (neighborhoods) around given points. The input is a cloud of n points in the plane without any user-defined parameters. The algorithm has a near linear time and a linear space O(n). The output is the array (number of holes, relative persistence in the filtration). We prove theoretical guarantees when the algorithm finds the correct number of holes (components in the complement) of an unknown shape approximated by a cloud.
****

Co-localization in Real-World Images
Kevin Tang, Armand Joulin, Li-Jia Li, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1464-1471
In this paper, we tackle the problem of co-localization in real-world images. Co-localization is the problem of simultaneously localizing (with bounding boxes) objects of the same class across a set of distinct images. Although similar problems such as co-segmentation and weakly supervised localization have been previously studied, we focus on being able to perform co-localization in real-world settings, which are typically characterized by large amounts of intra-class variation, inter-class diversity, and annotation noise. To address these issues, we present a joint image-box formulation for solving the co-localization problem, and show how it can be relaxed to a convex quadratic program which can be efficiently solved. We perform an extensive evaluation of our method compared to previous state-of-the-art approaches on the challenging PASCAL VOC 2007 and Object Discovery datasets. In addition, we also present a large-scale study of co-localization on ImageNet, involving ground-truth annotations for 3,624 classes and approximately 1 million images.
****

Spectral Clustering with Jensen-type Kernels and their Multi-point Extensions
Debarghya Ghoshdastidar, Ambedkar Dukkipati, Ajay P. Adsul, Aparna S. Vijayan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CV

PR), 2014, pp. 1472-1477

Motivated by multi-distribution divergences, which originate in information theory, we propose a notion of `multi-point' kernels, and study their applications. We study a class of kernels based on Jensen type divergences and show that these can be extended to measure similarity among multiple points. We study tensor flattening methods and develop a multi-point (kernel) spectral clustering (MSC) method. We further emphasize on a special case of the proposed kernels, which is a multi-point extension of the linear (dot-product) kernel and show the existence of cubic time tensor flattening algorithm in this case. Finally, we illustrate the usefulness of our contributions using standard data sets and image segmentation tasks.

****

Fast and Robust Archetypal Analysis for Representation Learning

Yuansi Chen, Julien Mairal, Zaid Harchaoui; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1478-1485

We revisit a pioneer unsupervised learning technique called archetypal analysis, which is related to successful data analysis methods such as sparse coding and non-negative matrix factorization. Since it was proposed, archetypal analysis did not gain a lot of popularity even though it produces more interpretable models than other alternatives. Because no efficient implementation has ever been made publicly available, its application to important scientific problems may have been severely limited. Our goal is to bring back into favour archetypal analysis. We propose a fast optimization scheme using an active-set strategy, and provide an efficient open-source implementation interfaced with Matlab, R, and Python. Then, we demonstrate the usefulness of archetypal analysis for computer vision tasks, such as codebook learning, signal classification, and large image collection visualization.

****

Photometric Bundle Adjustment for Dense Multi-View 3D Modeling

Amael Delaunoy, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1486-1493

Motivated by a Bayesian vision of the 3D multi-view reconstruction from images problem, we propose a dense 3D reconstruction technique that jointly refines the shape and the camera parameters of a scene by minimizing the photometric reprojection error between a generated model and the observed images, hence considering all pixels in the original images. The minimization is performed using a gradient descent scheme coherent with the shape representation (here a triangular mesh), where we derive evolution equations in order to optimize both the shape and the camera parameters. This can be used at a last refinement step in 3D reconstruction pipelines and helps improving the 3D reconstruction's quality by estimating the 3D shape and camera calibration more accurately. Examples are shown for multi-view stereo where the texture is also jointly optimized and improved, but could be used for any generative approaches dealing with multi-view reconstruction settings (i.e. depth map fusion, multi-view photometric stereo).

****

The Photometry of Intrinsic Images

Marc Serra, Olivier Penacchio, Robert Benavente, Maria Vanrell, Dimitris Samaras; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1494-1501

Intrinsic characterization of scenes is often the best way to overcome the illumination variability artifacts that complicate most computer vision problems, from 3D reconstruction to object or material recognition. This paper examines the deficiency of existing intrinsic image models to accurately account for the effects of illuminant color and sensor characteristics in the estimation of intrinsic images and presents a generic framework which incorporates insights from color constancy research to the intrinsic image decomposition problem. The proposed mathematical formulation includes information about the color of the illuminant and the effects of the camera sensors, both of which modify the observed color of the reflectance of the objects in the scene during the acquisition process. By modeling these effects, we get a "truly intrinsic" reflectance image, which w e call absolute reflectance, which is invariant to changes of illuminant or camera sensors. This model allows us to represent a wide range of intrinsic image decompositions depending on the specific assumptions on the geometric properties of the scene configuration and the spectral properties of the light source and the acquisition system, thus unifying previous models in a single general framework. We demonstrate that even partial information about sensors improves significantly the estimated reflectance images, thus making our method applicable for a wide range of sensors. We validate our general intrinsic image framework experimentally with both synthetic data and natural images.

****

High Resolution 3D Shape Texture from Multiple Videos

Vagia Tsiminaki, Jean-Sebastien Franco, Edmond Boyer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1502-1509

We examine the problem of retrieving high resolution textures of objects observed in multiple videos under small object deformations. In the monocular case, the data redundancy necessary to reconstruct a high-resolution image stems from temporal accumulation. This has been vastly explored and is known as image super-resolution. On the other hand, a handful of methods have considered the texture of a static 3D object observed from several cameras, where the data redundancy is obtained through the different viewpoints. We introduce a unified framework to leverage both possibilities for the estimation of an object's high resolution texture. This framework uniformly deals with any related geometric variability introduced by the acquisition chain or by the evolution over time. To this goal we use 2D warps for all viewpoints and all temporal frames and a linear image formation model from texture to image space. Despite its simplicity, the method is able to successfully handle different views over space and time. As shown experimentally, it demonstrates the interest of temporal information to improve the texture quality. Additionally, we also show that our method outperforms state of the art multi-view super-resolution methods existing for the static case.

****

PatchMatch Based Joint View Selection and Depthmap Estimation

Enliang Zheng, Enrique Dunn, Vladimir Jojic, Jan-Michael Frahm; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1510-1517

We propose a multi-view depthmap estimation approach aimed at adaptively ascertaining the pixel level data associations between a reference image and all the elements of a source image set. Namely, we address the question, what aggregation subset of the source image set should we use to estimate the depth of a particular pixel in the reference image? We pose the problem within a probabilistic framework that jointly models pixel-level view selection and depthmap estimation given the local pairwise image photoconsistency. The corresponding graphical model is solved by EM-based view selection probability inference and PatchMatch-like depth sampling and propagation. Experimental results on standard multi-view benchmarks convey the state-of-the art estimation accuracy afforded by mitigating spurious pixel level data associations. Additionally, experiments on large Internet crowd sourced data demonstrate the robustness of our approach against unstructured and heterogeneous image capture characteristics. Moreover, the linear computational and storage requirements of our formulation, as well as its inherent parallelism, enables an efficient and scalable GPU-based implementation.

****

Light Field Stereo Matching Using Bilateral Statistics of Surface Cameras

Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1518-1525

In this paper, we introduce a bilateral consistency metric on the surface camera (SCam) for light field stereo matching to handle significant occlusions. The concept of SCam is used to model angular radiance distribution with respect to a 3D point. Our bilateral consistency metric is used to indicate the probability of occlusions by analyzing the SCams. We further show how to distinguish between on-surface and free space, textured and non-textured, and Lambertian and specular

through bilateral SCam analysis. To speed up the matching process, we apply the edge preserving guided filter on the consistency-disparity curves. Experimental results show that our technique outperforms both the state-of-the-art and the recent light field stereo matching methods, especially near occlusion boundaries.
****

Recovering Surface Details under General Unknown Illumination Using Shading and Coarse Multi-view Stereo

Di Xu, Qi Duan, Jianming Zheng, Juyong Zhang, Jianfei Cai, Tat-Jen Cham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1526-1533

Reconstructing the shape of a 3D object from multi-view images under unknown, general illumination is a fundamental problem in computer vision and high quality reconstruction is usually challenging especially when high detail is needed. This paper presents a total variation (TV) based approach for recovering surface details using shading and multi-view stereo (MVS). Behind the approach are our two important observations: (1) the illumination over the surface of an object tends to be piecewise smooth and (2) the recovery of surface orientation is not sufficient for reconstructing geometry, which were previously overlooked. Thus we introduce TV to regularize the lighting and use visual hull to constrain partial vertices. The reconstruction is formulated as a constrained TVminimization problem that treats the shape and lighting as unknowns simultaneously. An augmented Lagrangian method is proposed to quickly solve the TV-minimization problem. As a result, our approach is robust, stable and is able to efficiently recover high quality of surface details even starting with a coarse MVS. These advantages are demonstrated by the experiments with synthetic and real world examples.
****

Probabilistic Labeling Cost for High-Accuracy Multi-View Reconstruction

Ilya Kostrikov, Esther Horbert, Bastian Leibe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1534-1541

In this paper, we propose a novel labeling cost for multi- view reconstruction. Existing approaches use data terms with specific weaknesses that are vulnerable to common challenges, such as low-textured regions or specularities. Our new probabilistic method implicitly discards outliers and can be shown to become more exact the closer we get to the true object surface. Our approach achieves top results among all published methods on the Middlebury DINO SPARSE dataset and also delivers accurate results on several other datasets with widely varying challenges, for which it works in unchanged form.
****

Complex Non-Rigid Motion 3D Reconstruction by Union of Subspaces

Yingying Zhu, Dong Huang, Fernando De La Torre, Simon Lucey; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1542-1549

The task of estimating complex non-rigid 3D motion through a monocular camera is of increasing interest to the wider scientific community. Assuming one has the 2D point tracks of the non-rigid object in question, the vision community refers to this problem as Non-Rigid Structure from Motion (NRSfM). In this paper we make two contributions. First, we demonstrate empirically that the current state of the art approach to NRSfM (i.e. Dai et al. [5]) exhibits poor reconstruction performance on complex motion (i.e motions involving a sequence of primitive actions such as walk, sit and stand involving a human object). Second, we propose that this limitation can be circumvented by modeling complex motion as a union of subspaces. This does not naturally occur in Dai et al.'s approach which instead makes a less compact summation of subspaces assumption. Experiments on both synthetic and real videos illustrate the benefits of our approach for the complex nonrigid motion analysis.
****

A Procrustean Markov Process for Non-Rigid Structure Recovery

Minsik Lee, Chong-Ho Choi, Songhwai Oh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1550-1557

Recovering a non-rigid 3D structure from a series of 2D observations is still a difficult problem to solve accurately. Many constraints have been proposed to facilitate the recovery, and one of the most successful constraints is smoothness due to the fact that most real-world objects change continuously. However, many existing methods require to determine the degree of smoothness beforehand, which is not viable in practical situations. In this paper, we propose a new probabilistic model that incorporates the smoothness constraint without requiring any prior knowledge. Our approach regards the sequence of 3D shapes as a simple stationary Markov process with Procrustes alignment, whose parameters are learned during the fitting process. The Markov process is assumed to be stationary because deformation is finite and recurrent in general, and the 3D shapes are assumed to be Procrustes aligned in order to discriminate deformation from motion. The proposed method outperforms the state-of-the-art methods, even though the computation time is rather moderate compared to the other existing methods.

****

Good Vibrations: A Modal Analysis Approach for Sequential Non-Rigid Structure from Motion

Antonio Agudo, Lourdes Agapito, Begona Calvo, Jose M. M. Montiel; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1558-1565

We propose an online solution to non-rigid structure from motion that performs camera pose and 3D shape estimation of highly deformable surfaces on a frame-by-frame basis. Our method models non-rigid deformations as a linear combination of some mode shapes obtained using modal analysis from continuum mechanics. The shape is first discretized into linear elastic triangles, modelled by means of finite elements, which are used to pose the force balance equations for an undamped free vibrations model. The shape basis computation comes down to solving an eigenvalue problem, without the requirement of a learning step. The camera pose and time varying weights that define the shape at each frame are then estimated on the fly, in an online fashion, using bundle adjustment over a sliding window of image frames. The result is a low computational cost method that can run sequentially in real-time. We show experimental results on synthetic sequences with ground truth 3D data and real videos for different scenarios ranging from sparse to dense scenes. Our system exhibits a good trade-off between accuracy and computational budget, it can handle missing data and performs favourably compared to competing methods.

****

Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving

Shiyu Song, Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1566-1573

Scale drift is a crucial challenge for monocular autonomous driving to emulate the performance of stereo. This paper presents a real-time monocular SFM system that corrects for scale drift using a novel cue combination framework for ground plane estimation, yielding accuracy comparable to stereo over long driving sequences. Our ground plane estimation uses multiple cues like sparse features, dense inter-frame stereo and (when applicable) object detection. A data-driven mechanism is proposed to learn models from training data that relate observation covariances for each cue to error behavior of its underlying variables. During testing, this allows per-frame adaptation of observation covariances based on relative confidences inferred from visual data. Our framework significantly boosts not only the accuracy of monocular self-localization, but also that of applications like object localization that rely on the ground plane. Experiments on the KITTI dataset demonstrate the accuracy of our ground plane estimation, monocular SFM and object localization relative to ground truth, with detailed comparisons to prior art.

****

On the Quotient Representation for the Essential Manifold

Roberto Tron, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1574-1581

The essential matrix, which encodes the epipolar constraint between points in two projective views, is a cornerstone of modern computer vision. Previous works h

ave proposed different characterizations of the space of essential matrices as a Riemannian manifold. However, they either do not consider the symmetric role played by the two views, or do not fully take into account the geometric peculiarities of the epipolar constraint. We address these limitations with a characterization as a quotient manifold which can be easily interpreted in terms of camera poses. While our main focus in on theoretical aspects, we include experiments in pose averaging, and show that the proposed formulation produces a meaningful distance between essential matrices.
****

Efficient High-Resolution Stereo Matching using Local Plane Sweeps
Sudipta N. Sinha, Daniel Scharstein, Richard Szeliski; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1582-1589
We present a stereo algorithm designed for speed and efficiency that uses local slanted plane sweeps to propose disparity hypotheses for a semi-global matching algorithm. Our local plane hypotheses are derived from initial sparse feature correspondences followed by an iterative clustering step. Local plane sweeps are then performed around each slanted plane to produce out-of-plane parallax and matching-cost estimates. A final global optimization stage, implemented using semi-global matching, assigns each pixel to one of the local plane hypotheses. By only exploring a small fraction of the whole disparity space volume, our technique achieves significant speedups over previous algorithms and achieves state-of-the-art accuracy on high-resolution stereo pairs of up to 19 megapixels.
****

Cross-Scale Cost Aggregation for Stereo Matching
Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, Shuicheng Yan, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1590-1597
Human beings process stereoscopic correspondence across multiple scales. However, this bio-inspiration is ignored by state-of-the-art cost aggregation methods for dense stereo correspondence. In this paper, a generic cross-scale cost aggregation framework is proposed to allow multi-scale interaction in cost aggregation. We firstly reformulate cost aggregation from a unified optimization perspective and show that different cost aggregation methods essentially differ in the choices of similarity kernels. Then, an inter-scale regularizer is introduced into optimization and solving this new optimization problem leads to the proposed framework. Since the regularization term is independent of the similarity kernel, various cost aggregation methods can be integrated into the proposed general framework. We show that the cross-scale framework is important as it effectively and efficiently expands state-of-the-art cost aggregation methods and leads to significant improvements, when evaluated on Middlebury, KITTI and New Tsukuba datasets.
****

Asymmetrical Gauss Mixture Models for Point Sets Matching
Wenbing Tao, Kun Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1598-1605
The probabilistic methods based on Symmetrical Gauss Mixture Model (SGMM) have achieved great success in point sets registration, but are seldom used to find the correspondences between two images due to the complexity of the non-rigid transformation and too many outliers. In this paper we propose an Asymmetrical GMM (AGMM) for point sets matching between a pair of images. Different from the previous SGMM, the AGMM gives each Gauss component a different weight which is related to the feature similarity between the data point and model point, which leads to two effective algorithms: the Single Gauss Model for Mismatch Rejection (SGMR) algorithm and the AGMM algorithm for point sets matching. The SGMR algorithm iteratively filters mismatches by estimating a non-rigid transformation between two images based on the spatial coherence of point sets. The AGMM algorithm combines the feature information with position information of the SIFT feature points extracted from the images to achieve point sets matching so that much more correct correspondences with high precision can be found. A number of comparison and evaluation experiments reveal the excellent performance of the proposed SGMR al

gorithm and AGMM algorithm.

****

Fast and Reliable Two-View Translation Estimation

Johan Fredriksson, Olof Enqvist, Fredrik Kahl; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1606-1612

It has long been recognized that one of the fundamental difficulties in theestimation of two-view epipolar geometry is the capability of handling outliers. In this paper, we develop a fast and tractable algorithm that maximizes the number of inliers under the assumption of a purely translating camera. Compared to classical random sampling methods, our approach is guaranteed to compute the optimal solution of a cost function based on reprojection errors and it has better time complexity. The performance is in fact independent of the inlier/outlier ratio of the data.This opens up for a more reliable approach to robust ego-motion estimation. Our basic translation estimator can be embedded into a system that computes the full camera rotation. We demonstrate the applicability in several difficult settings with large amounts of outliers. It turns out to be particularly well-suited for small rotations and rotations around a known axis (which is the case for cellular phones where the gravitation axis can be measured). Experimental results show that compared to standard ransac methods based on minimal solvers, ouralgorithm produces more accurate estimates in the presence of large outlier ratios.

****

Graph Cut based Continuous Stereo Matching using Locally Shared Labels

Tatsunori Taniai, Yasuyuki Matsushita, Takeshi Naemura; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1613-1620

We present an accurate and efficient stereo matching method using locally shared labels, a new labeling scheme that enables spatial propagation in MRF inference using graph cuts. They give each pixel and region a set of candidate disparity labels, which are randomly initialized, spatially propagated, and refined for continuous disparity estimation. We cast the selection and propagation of locallydefined disparity labels as fusion-based energy minimization. The joint use of graph cuts and locally shared labels has advantages over previous approaches based on fusion moves or belief propagation; it produces submodular moves deriving a subproblem optimality; enables powerful randomized search; helps to find good smooth, locally planar disparity maps, which are reasonable for natural scenes; allows parallel computation of both unary and pairwise costs. Our method is evaluated using the Middlebury stereo benchmark and achieves first place in sub-pixel accuracy.

****

Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching

Aristotle Spyropoulos, Nikos Komodakis, Philippos Mordohai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1621-1628

While machine learning has been instrumental to the ongoing progress in most areas of computer vision, it has not been applied to the problem of stereo matching with similar frequency or success. We present a supervised learning approach for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel. We show highly competitive results in predicting the correctness of matches and in confidence estimation, which allows us to rank pixels according to the reliability of their assigned disparities. Moreover, we show how these confidence values can be used to improve the accuracy of disparity maps by integrating them with an MRF-based stereo algorithm. This is an important distinction from current literature that has mainly focused on sparsification by removing potentially erroneous disparities to generate quasi-dense disparity maps.

****

Decorrelating Semantic Visual Attributes by Resisting the Urge to Share

Dinesh Jayaraman, Fei Sha, Kristen Grauman; Proceedings of the IEEE Conference o n Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1629-1636

Existing methods to learn visual attributes are prone to learning the wrong thing---namely, properties that are correlated with the attribute of interest among training samples. Yet, many proposed applications of attributes rely on being able to learn the correct semantic concept corresponding to each attribute. We propose to resolve such confusions by jointly learning decorrelated, discriminative attribute models. Leveraging side information about semantic relatedness, we develop a multi-task learning approach that uses structured sparsity to encourage feature competition among unrelated attributes and feature sharing among related attributes. On three challenging datasets, we show that accounting for structure in the visual attribute space is key to learning attribute models that preserve semantics, yielding improved generalizability that helps in the recognition and discovery of unseen object categories.

****

PANDA: Pose Aligned Networks for Deep Attribute Modeling

Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, Lubomir Bourdev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1637-1644

We propose a method for inferring human attributes (such as gender, hair style, clothes style, expression, action) from images of people under large variation of viewpoint, pose, appearance, articulation and occlusion. Convolutional Neural Nets (CNN) have been shown to perform very well on large scale object recognition problems. In the context of attribute classification, however, the signal is often subtle and it may cover only a small part of the image, while the image is dominated by the effects of pose and viewpoint. Discounting for pose variation would require training on very large labeled datasets which are not presently available. Part-based models, such as poselets and DPM have been shown to perform well for this problem but they are limited by shallow low-level features. We propose a new method which combines part-based models and deep learning by training pose-normalized CNNs. We show substantial improvement vs. state-of-the-art methods on challenging attribute classification tasks in unconstrained settings. Experiments confirm that our method outperforms both the best part-based methods on this problem and conventional CNNs trained on the full bounding box of the person.

****

Learning Scalable Discriminative Dictionary with Sample Relatedness

Jiashi Feng, Stefanie Jegelka, Shuicheng Yan, Trevor Darrell; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1645-1652

Attributes are widely used as mid-level descriptors of object properties in object recognition and retrieval. Mostly, such attributes are manually pre-defined based on domain knowledge, and their number is fixed. However, pre-defined attributes may fail to adapt to the properties of the data at hand, may not necessarily be discriminative, and/or may not generalize well. In this work, we propose a dictionary learning framework that flexibly adapts to the complexity of the given data set and reliably discovers the inherent discriminative middle-level binary features in the data. We use sample relatedness information to improve the generalization of the learned dictionary. We demonstrate that our framework is applicable to both object recognition and complex image retrieval tasks even with few training examples. Moreover, the learned dictionary also help classify novel object categories. Experimental results on the Animals with Attributes, ILSVRC2010 and PASCAL VOC2007 datasets indicate that using relatedness information leads to significant performance gains over established baselines.

****

DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev, Christian Szegedy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653-1660

We propose a method for human pose estimation based on Deep Neural Networks (DNNs). The pose estimation is formulated as a DNN-based regression problem towards body joints. We present a cascade of such DNN regressors which results in high p recision pose estimates. The approach has the advantage of reasoning about pose in a holistic fashion and has a simple but yet powerful formulation which capitalizes on recent advances in Deep Learning. We present a detailed empirical analysis with state-of-art or better performance on four academic benchmarks of diverse real-world images.

****

Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation

Catalin Ionescu, Joao Carreira, Cristian Sminchisescu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1661-1668

Recently, the emergence of Kinect systems has demonstrated the benefits of predicting an intermediate body part labeling for 3D human pose estimation, in conjunction with RGB-D imagery. The availability of depth information plays a critical role, so an important question is whether a similar representation can be developed with sufficient robustness in order to estimate 3D pose from RGB images. This paper provides evidence for a positive answer, by leveraging (a) 2D human body part labeling in images, (b) second-order label-sensitive pooling over dynamically computed regions resulting from a hierarchical decomposition of the body, and (c) iterative structured-output modeling to contextualize the process based on 3D pose estimates. For robustness and generalization, we take advantage of a recent large-scale 3D human motion capture dataset, Human3.6M [18] that also has human body part labeling annotations available with images. We provide extensive experimental studies where alternative intermediate representations are compared and report a substantial 33% error reduction over competitive discriminative baselines that regress 3D human pose against global HOG features.

****

3D Pictorial Structures for Multiple Human Pose Estimation

Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, Slobodan Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1669-1676

In this work, we address the problem of 3D pose estimation of multiple humans from multiple views. This is a more challenging problem than single human 3D pose estimation due to the much larger state space, partial occlusions as well as across view ambiguities when not knowing the identity of the humans in advance. To address these problems, we first create a reduced state space by triangulation of corresponding body joints obtained from part detectors in pairs of camera views. In order to resolve the ambiguities of wrong and mixed body parts of multiple humans after triangulation and also those coming from false positive body part detections, we introduce a novel 3D pictorial structures (3DPS) model. Our model infers 3D human body configurations from our reduced state space. The 3DPS model is generic and applicable to both single and multiple human pose estimation. In order to compare to the state-of-the art, we first evaluate our method on single human 3D pose estimation on HumanEva-I [22] and KTH Multiview Football Dataset II [8] datasets. Then, we introduce and evaluate our method on two datasets for multiple human 3D pose estimation.

****

Learning Euclidean-to-Riemannian Metric for Point-to-Set Classification

Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1677-1684

In this paper, we focus on the problem of point-to-set classification, where single points are matched against sets of correlated points. Since the points commonly lie in Euclidean space while the sets are typically modeled as elements on Riemannian manifold, they can be treated as Euclidean points and Riemannian points respectively. To learn a metric between the heterogeneous points, we propose a novel Euclidean-to-Riemannian metric learning framework. Specifically, by exploiting typical Riemannian metrics, the Riemannian manifold is first embedded into a high dimensional Hilbert space to reduce the gaps between the heterogeneous spaces and meanwhile respect the Riemannian geometry of the manifold. The final distance metric is then learned by pursuing multiple transformations from the Hilbert space and the original Euclidean space (or its corresponding Hilbert space) to a common Euclidean subspace, where classical Euclidean distances of transfor med heterogeneous points can be measured. Extensive experiments clearly demonstrate the superiority of our proposed approach over the state-of-the-art methods.

****

Face Alignment at 3000 FPS via Regressing Local Binary Features

Shaoqing Ren, Xudong Cao, Yichen Wei, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1685-1692

This paper presents a highly efficient, very accurate regression approach for face alignment. Our approach has two novel components: a set of local binary features, and a locality principle for learning those features. The locality principle guides us to learn a set of highly discriminative local binary features for each facial landmark independently. The obtained local binary features are used to jointly learn a linear regression for the final output. Our approach achieves the state-of-the-art results when tested on the current most challenging benchmarks. Furthermore, because extracting and regressing local binary features is computationally very cheap, our system is much faster than previous methods. It achieves over 3,000 fps on a desktop or 300 fps on a mobile phone for locating a few dozens of landmarks.

****

A Compact and Discriminative Face Track Descriptor

Omkar M. Parkhi, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1693-1700

Our goal is to learn a compact, discriminative vector representation of a face track, suitable for the face recognition tasks of verification and classification. To this end, we propose a novel face track descriptor, based on the Fisher Vector representation, and demonstrate that it has a number of favourable properties. First, the descriptor is suitable for tracks of both frontal and profile faces, and is insensitive to their pose. Second, the descriptor is compact due to discriminative dimensionality reduction, and it can be further compressed using binarization. Third, the descriptor can be computed quickly (using hard quantization) and its compact size and fast computation render it very suitable for large scale visual repositories. Finally, the descriptor demonstrates good generalization when trained on one dataset and tested on another, reflecting its tolerance to the dataset bias. In the experiments we show that the descriptor exceeds the state of the art on both face verification task (YouTube Faces without outside training data, and INRIA-Buffy benchmarks), and face classification task (using the Oxford-Buffy dataset).

****

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701-1708

In modern face recognition, the conventional pipeline consists of four stages: detect => align => represent => classify. We revisit both the alignment step and the representation step by employing explicit 3D face modeling in order to apply a piecewise affine transformation, and derive a face representation from a nine-layer deep neural network. This deep network involves more than 120 million parameters using several locally connected layers without weight sharing, rather than the standard convolutional layers. Thus we trained it on the largest facial dataset to-date, an identity labeled dataset of four million facial images belonging to more than 4,000 identities. The learned representations coupling the accurate model-based alignment with the large facial database generalize remarkably well to faces in unconstrained environments, even with a simple classifier. Our method reaches an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset, reducing the error of the current state of the art by more than 27%, closely approaching human-level performance.

****

Filter Forests for Learning Data-Dependent Convolutional Kernels

Sean Ryan Fanello, Cem Keskin, Pushmeet Kohli, Shahram Izadi, Jamie Shotton, Antonio Criminisi, Ugo Pattacini, Tim Paek; Proceedings of the IEEE Conference on C omputer Vision and Pattern Recognition (CVPR), 2014, pp. 1709-1716

We propose 'filter forests' (FF), an efficient new discriminative approach for predicting continuous variables given a signal and its context. FF can be used for general signal restoration tasks that can be tackled via convolutional filtering, where it attempts to learn the optimal filtering kernels to be applied to each data point. The model can learn both the size of the kernel and its values, conditioned on the observation and its spatial or temporal context. We show that FF compares favorably to both Markov random field based and recently proposed regression forest based approaches for labeling problems in terms of efficiency and accuracy. In particular, we demonstrate how FF can be used to learn optimal denoising filters for natural images as well as for other tasks such as depth image refinement, and 1D signal magnitude estimation. Numerous experiments and quantitative comparisons show that FFs achieve accuracy at par or superior to recent state of the art techniques, while being several orders of magnitude faster.

****

Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks

Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1717-1724

Convolutional neural networks (CNN) have recently shown outstanding image classification performance in the large- scale visual recognition challenge (ILSVRC2012). The success of CNNs is attributed to their ability to learn rich mid-level image representations as opposed to hand-designed low-level features used in other image classification methods. Learning CNNs, however, amounts to estimating millions of parameters and requires a very large number of annotated image samples. This property currently prevents application of CNNs to problems with limited training data. In this work we show how image representations learned with CNNs on large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data. We design a method to reuse layers trained on the ImageNet dataset to compute mid-level image representation for images in the PASCAL VOC dataset. We show that despite differences in image statistics and tasks in the two datasets, the transferred representation leads to significantly improved results for object and action classification, outperforming the current state of the art on Pascal VOC 2007 and 2012 datasets. We also show promising results for object and action localization.

****

Large-scale Video Classification with Convolutional Neural Networks

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725-1732

Convolutional Neural Networks (CNNs) have been established as a powerful class of models for image recognition problems. Encouraged by these results, we provide an extensive empirical evaluation of CNNs on large-scale video classification using a new dataset of 1 million YouTube videos belonging to 487 classes. We study multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggest a multiresolution, foveated architecture as a promising way of speeding up the training. Our best spatio-temporal networks display significant performance improvements compared to strong feature-based baselines (55.3% to 63.9%), but only a surprisingly modest improvement compared to single-frame models (59.3% to 60.9%). We further study the generalization performance of our best model by retraining the top layers on the UCF-101 Action Recognition dataset and observe significant performance improvements compared to the UCF-101 baseline model (63.3% up from 43.9%).

****

Convolutional Neural Networks for No-Reference Image Quality Assessment

Le Kang, Peng Ye, Yi Li, David Doermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1733-1740

In this work we describe a Convolutional Neural Network (CNN) to accurately predict image quality without a reference image. Taking image patches as input, the CNN works in the spatial domain without using hand-crafted features that are emp loyed by most previous methods. The network consists of one convolutional layer with max and min pooling, two fully connected layers and an output node. Within the network structure, feature learning and regression are integrated into one optimization process, which leads to a more effective model for estimating image quality. This approach achieves state of the art performance on the LIVE dataset and shows excellent generalization ability in cross dataset experiments. Further experiments on images with local distortions demonstrate the local quality estimation ability of our CNN, which is rarely reported in previous literature.

****

Nonparametric Context Modeling of Local Appearance for Pose- and Expression-Robust Facial Landmark Localization

Brandon M. Smith, Jonathan Brandt, Zhe Lin, Li Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1741-1748

We propose a data-driven approach to facial landmark localization that models the correlations between each landmark and its surrounding appearance features. At runtime, each feature casts a weighted vote to predict landmark locations, where the weight is precomputed to take into account the feature's discriminative power. The feature votingbased landmark detection is more robust than previous local appearance-based detectors; we combine it with nonparametric shape regularization to build a novel facial landmark localization pipeline that is robust to scale, in-plane rotation, occlusion, expression, and most importantly, extreme head pose. We achieve state-of-the-art performance on two especially challenging in-the-wild datasets populated by faces with extreme head pose and expression.

****

Learning Expressionlets on Spatio-Temporal Manifold for Dynamic Facial Expression Recognition

Mengyi Liu, Shiguang Shan, Ruiping Wang, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1749-1756

Facial expression is temporally dynamic event which can be decomposed into a set of muscle motions occurring in different facial regions over various time intervals. For dynamic expression recognition, two key issues, temporal alignment and semantics-aware dynamic representation, must be taken into account. In this paper, we attempt to solve both problems via manifold modeling of videos based on a novel mid-level representation, i.e. expressionlet. Specifically, our method contains three key components: 1) each expression video clip is modeled as a spatio-temporal manifold (STM) formed by dense low-level features; 2) a Universal Manifold Model (UMM) is learned over all low-level features and represented as a set of local ST modes to statistically unify all the STMs. 3) the local modes on each STM can be instantiated by fitting to UMM, and the corresponding expressionlet is constructed by modeling the variations in each local ST mode. With above strategy, expression videos are naturally aligned both spatially and temporally. To enhance the discriminative power, the expressionlet-based STM representation is further processed with discriminant embedding. Our method is evaluated on four public expression databases, CK+, MMI, Oulu-CASIA, and AFEW. In all cases, our method reports results better than the known state-of-the-art.

****

Who Do I Look Like? Determining Parent-Offspring Resemblance via Gated Autoencoders

Afshin Dehghan, Enrique G. Ortiz, Ruben Villegas, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1757-1764

Recent years have seen a major push for face recognition technology due to the large expansion of image sharing on social networks. In this paper, we consider the difficult task of determining parent-offspring resemblance using deep learning to answer the question "Who do I look like?" Although humans can perform this job at a rate higher than chance, it is not clear how they do it [2]. However, recent studies in anthropology [24] have determined which features tend to be the most discriminative. In this study, we aim to not only create an accurate system for resemblance detection, but bridge the gap between studies in anthropology with computer vision techniques. Further, we aim to answer two key questions: 1)

Do offspring resemble their parents? and 2) Do offspring resemble one parent more than the other? We propose an algorithm that fuses the features and metrics discovered via gated autoencoders with a discriminative neural network layer that learns the optimal, or what we call genetic, features to delineate parent-offspring relationships. We further analyze the correlation between our automatically detected features and those found in anthropological studies. Meanwhile, our method outperforms the state-of-the-art in kinship verification by 3-10% depending on the relationship using specific (father-son, mother-daughter, etc.) and generic models.

****

Unified Face Analysis by Iterative Multi-Output Random Forests

Xiaowei Zhao, Tae-Kyun Kim, Wenhan Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1765-1772

In this paper, we present a unified method for joint face image analysis, i.e., simultaneously estimating head pose, facial expression and landmark positions in real-world face images. To achieve this goal, we propose a novel iterative Multi-Output Random Forests (iMORF) algorithm, which explicitly models the relations among multiple tasks and iteratively exploits such relations to boost the performance of all tasks. Specifically, a hierarchical face analysis forest is learned to perform classification of pose and expression at the top level, while performing landmark positions regression at the bottom level. On one hand, the estimated pose and expression provide strong shape prior to constrain the variation of landmark positions. On the other hand, more discriminative shape-related features could be extracted from the estimated landmark positions to further improve the predictions of pose and expression. This relatedness of face analysis tasks is iteratively exploited through several cascaded hierarchical face analysis forests until convergence. Experiments conducted on publicly available real-world face datasets demonstrate that the performance of all individual tasks are significantly improved by the proposed iMORF algorithm. In addition, our method outperforms state-of-the-arts for all three face analysis tasks.

****

Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras

Kenneth Alberto Funes Mora, Jean-Marc Odobez; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1773-1780

We propose a head pose invariant gaze estimation model for distant RGB-D cameras. It relies on a geometric understanding of the 3D gaze action and generation of eye images. By introducing a semantic segmentation of the eye region within a generative process, the model (i) avoids the critical feature tracking of geometrical approaches requiring high resolution images; (ii) decouples the person dependent geometry from the ambient conditions, allowing adaptation to different conditions without retraining. Priors in the generative framework are adequate for training from few samples. In addition, the model is capable of gaze extrapolation allowing for less restrictive training schemes. Comparisons with state of the art methods validate these properties which make our method highly valuable for addressing many diverse tasks in sociology, HRI and HCI.

****

A Hierarchical Probabilistic Model for Facial Feature Detection

Yue Wu, Ziheng Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1781-1788

Facial feature detection from facial images has attracted great attention in the field of computer vision. It is a nontrivial task since the appearance and shape of the face tend to change under different conditions. In this paper, we propose a hierarchical probabilistic model that could infer the true locations of facial features given the image measurements even if the face is with significant facial expression and pose. The hierarchical model implicitly captures the lower level shape variations of facial components using the mixture model. Furthermore, in the higher level, it also learns the joint relationship among facial components, the facial expression, and the pose information through automatic structure learning and parameter estimation of the probabilistic model. Experimental results on benchmark databases demonstrate the effectiveness of the proposed hierar chical probabilistic model.

****

RAPS: Robust and Efficient Automatic Construction of Person-Specific Deformable Models

Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1789-1796

The construction of Facial Deformable Models (FDMs) is a very challenging computer vision problem, since the face is a highly deformable object and its appearance drastically changes under different poses, expressions, and illuminations. Although several methods for generic FDMs construction, have been proposed for facial landmark localization in still images, they are insufficient for tasks such as facial behaviour analysis and facial motion capture where perfect landmark localization is required. In this case, person-specific FDMs (PSMs) are mainly employed, requiring manual facial landmark annotation for each person and person-specific training. In this paper, a novel method for the automatic construction of PSMs is proposed. To this end, an orthonormal subspace which is suitable for facial image reconstruction is learnt. Next, to correct the fittings of a generic model, image congealing (i.e., batch image aliment) is performed by employing only the learnt orthonormal subspace. Finally, the corrected fittings are used to construct the PSM. The image congealing problem is solved by formulating a suitable sparsity regularized rank minimization problem. The proposed method outperforms the state-of-the-art methods that is compared to, in terms of both landmark localization accuracy and computational time.

****

Non-Parametric Bayesian Constrained Local Models

Pedro Martins, Rui Caseiro, Jorge Batista; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1797-1804

This work presents a novel non-parametric Bayesian formulation for aligning faces in unseen images. Popular approaches, such as the Constrained Local Models (CLM) or the Active Shape Models (ASM), perform facial alignment through a local search, combining an ensemble of detectors with a global optimization strategy that constraints the facial feature points to be within the subspace spanned by a Point Distribution Model (PDM). The global optimization can be posed as a Bayesian inference problem, looking to maximize the posterior distribution of the PDM parameters in a maximum a posteriori (MAP) sense. Previous approaches rely exclusively on Gaussian inference techniques, i.e. both the likelihood (detectors responses) and the prior (PDM) are Gaussians, resulting in a posterior which is also Gaussian, whereas in this work the posterior distribution is modeled as being non-parametric by a Kernel Density Estimator (KDE). We show that this posterior distribution can be efficiently inferred using Sequential Monte Carlo methods, in particular using a Regularized Particle Filter (RPF). The technique is evaluated in detail on several standard datasets (IMM, BioID, XM2VTS, LFW and FGNET Talking Face) and compared against state-of-the-art CLM methods. We demonstrate that inferring the PDM parameters non-parametrically significantly increase the face alignment performance.

****

Facial Expression Recognition via a Boosted Deep Belief Network

Ping Liu, Shizhong Han, Zibo Meng, Yan Tong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1805-1812

A training process for facial expression recognition is usually performed sequentially in three individual stages: feature learning, feature selection, and classifier construction. Extensive empirical studies are needed to search for an optimal combination of feature representation, feature set, and classifier to achieve good recognition performance. This paper presents a novel Boosted Deep Belief Network (BDBN) for performing the three training stages iteratively in a unified loopy framework. Through the proposed BDBN framework, a set of features, which is effective to characterize expression-related facial appearance/shape changes, can be learned and selected to form a boosted strong classifier in a statistical way. As learning continues, the strong classifier is improved iteratively an d more importantly, the discriminative capabilities of selected features are strengthened as well according to their relative importance to the strong classifier via a joint fine-tune process in the BDBN framework. Extensive experiments on two public databases showed that the BDBN framework yielded dramatic improvements in facial expression analysis.

****

Automatic Construction of Deformable Models In-The-Wild

Epameinondas Antonakos, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1813-1820

Deformable objects are everywhere. Faces, cars, bicycles, chairs etc. Recently, there has been a wealth of research on training deformable models for object detection, part localization and recognition using annotated data. In order to train deformable models with good generalization ability, a large amount of carefully annotated data is required, which is a highly time consuming and costly task. We propose the first - to the best of our knowledge - method for automatic construction of deformable models using images captured in totally unconstrained conditions, recently referred to as "in-the-wild". The only requirements of the method are a crude bounding box object detector and a-priori knowledge of the object's shape (e.g. a point distribution model). The object detector can be as simple as the Viola-Jones algorithm (e.g. even the cheapest digital camera features a robust face detector). The 2D shape model can be created by using only a few shape examples with deformations. In our experiments on facial deformable models, we show that the proposed automatically built model not only performs well, but also outperforms discriminative models trained on carefully annotated data. To the best of our knowledge, this is the first time it is shown that an automatically constructed model can perform as well as methods trained directly on annotated data.

****

Learning-by-Synthesis for Appearance-based 3D Gaze Estimation

Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1821-1828

Inferring human gaze from low-resolution eye images is still a challenging task despite its practical importance in many application scenarios. This paper presents a learning-by-synthesis approach to accurate image-based gaze estimation that is person- and head pose-independent. Unlike existing appearance-based methods that assume person-specific training data, we use a large amount of cross-subject training data to train a 3D gaze estimator. We collect the largest and fully calibrated multi-view gaze dataset and perform a 3D reconstruction in order to generate dense training data of eye images. By using the synthesized dataset to learn a random regression forest, we show that our method outperforms existing methods that use low-resolution eye images.

****

Towards Multi-view and Partially-Occluded Face Alignment

Junliang Xing, Zhiheng Niu, Junshi Huang, Weiming Hu, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1829-1836

We present a robust model to locate facial landmarks under different views and possibly severe occlusions. To build reliable relationships between face appearance and shape with large view variations, we propose to formulate face alignment as an L1-induced Stagewise Relational Dictionary (SRD) learning problem. During each training stage, the SRD model learns a relational dictionary to capture consistent relationships between face appearance and shape, which are respectively modeled by the pose-indexed image features and the shape displacements for current estimated landmarks. During testing, the SRD model automatically selects a sparse set of the most related shape displacements for the testing face and uses them to refine its shape iteratively. To locate facial landmarks under occlusions, we further propose to learn an occlusion dictionary to model different kinds of partial face occlusions. By deploying the occlusion dictionary into the SRD model, the alignment performance for occluded faces can be further improved. Our algorithm is simple, effective, and easy to implement. Extensive experiments on t wo benchmark datasets and two newly built datasets have demonstrated its superior performances over the state-of-the-art methods, especially for faces with large view variations and/or occlusions.

****

Head Pose Estimation Based on Multivariate Label Distribution

Xin Geng, Yu Xia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1837-1842

Accurate ground truth pose is essential to the training of most existing head pose estimation algorithms. However, in many cases, the "ground truth" pose is obtained in rather subjective ways, such as asking the human subjects to stare at different markers on the wall. In such case, it is better to use soft labels rather than explicit hard labels. Therefore, this paper proposes to associate a multivariate label distribution (MLD) to each image. An MLD covers a neighborhood around the original pose. Labeling the images with MLD can not only alleviate the problem of inaccurate pose labels, but also boost the training examples associated to each pose without actually increasing the total amount of training examples. Two algorithms are proposed to learn from the MLD by minimizing the weighted Jeffrey's divergence between the predicted MLD and the ground truth MLD. Experimental results show that the MLD-based methods perform significantly better than the compared state-of-the-art head pose estimation algorithms.

****

Efficient Boosted Exemplar-based Face Detection

Haoxiang Li, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Gang Hua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1843-1850

Despite the fact that face detection has been studied intensively over the past several decades, the problem is still not completely solved. Challenging conditions, such as extreme pose, lighting, and occlusion, have historically hampered traditional, model-based methods. In contrast, exemplar-based face detection has been shown to be effective, even under these challenging conditions, primarily because a large exemplar database is leveraged to cover all possible visual variations. However, relying heavily on a large exemplar database to deal with the face appearance variations makes the detector impractical due to the high space and time complexity. We construct an efficient boosted exemplar-based face detector which overcomes the defect of the previous work by being faster, more memory efficient, and more accurate. In our method, exemplars as weak detectors are discriminatively trained and selectively assembled in the boosting framework which largely reduces the number of required exemplars. Notably, we propose to include non-face images as negative exemplars to actively suppress false detections to further improve the detection accuracy. We verify our approach over two public face detection benchmarks and one personal photo album, and achieve significant improvement over the state-of-the-art algorithms in terms of both accuracy and efficiency.

****

Gauss-Newton Deformable Part Models for Face Alignment in-the-Wild

Georgios Tzimiropoulos, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1851-1858

Arguably, Deformable Part Models (DPMs) are one of the most prominent approaches for face alignment with impressive results being recently reported for both controlled lab and unconstrained settings. Fitting in most DPM methods is typically formulated as a two-step process during which discriminatively trained part templates are first correlated with the image to yield a filter response for each landmark and then shape optimization is performed over these filter responses. This process, although computationally efficient, is based on fixed part templates which are assumed to be independent, and has been shown to result in imperfect filter responses and detection ambiguities. To address this limitation, in this paper, we propose to jointly optimize a part-based, trained in-the-wild, flexible appearance model along with a global shape model which results in a joint translational motion model for the model parts via Gauss-Newton (GN) optimization. We show how significant computational reductions can be achieved by building a fu ll model during training but then efficiently optimizing the proposed cost function on a sparse grid using weighted least-squares during fitting. We coin the proposed formulation Gauss-Newton Deformable Part Model (GN-DPM). Finally, we compare its performance against the state-of-the-art and show that the proposed GN-DPM outperforms it, in some cases, by a large margin. Code for our method is available from http://ibug.doc.ic.ac.uk/resources

****

Incremental Face Alignment in the Wild

Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, Maja Pantic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1859-1866

The development of facial databases with an abundance of annotated facial data captured under unconstrained 'in-the-wild' conditions have made discriminative facial deformable models the de facto choice for generic facial landmark localization. Even though very good performance for the facial landmark localization has been shown by many recently proposed discriminative techniques, when it comes to the applications that require excellent accuracy, such as facial behaviour analysis and facial motion capture, the semi-automatic person-specific or even tedious manual tracking is still the preferred choice. One way to construct a person-specific model automatically is through incremental updating of the generic model. This paper deals with the problem of updating a discriminative facial deformable model, a problem that has not been thoroughly studied in the literature. In particular, we study for the first time, to the best of our knowledge, the strategies to update a discriminative model that is trained by a cascade of regressors. We propose very efficient strategies to update the model and we show that is possible to automatically construct robust discriminative person and imaging condition specific models 'in-the-wild' that outperform state-of-the-art generic face alignment strategies.

****

One Millisecond Face Alignment with an Ensemble of Regression Trees

Vahid Kazemi, Josephine Sullivan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867-1874

This paper addresses the problem of Face Alignment for a single image. We show how an ensemble of regression trees can be used to estimate the face's landmark positions directly from a sparse subset of pixel intensities, achieving super-realtime performance with high quality predictions. We present a general framework based on gradient boosting for learning an ensemble of regression trees that optimizes the sum of square error loss and naturally handles missing or partially labelled data. We show how using appropriate priors exploiting the structure of image data helps with efficient feature selection. Different regularization strategies and its importance to combat overfitting are also investigated. In addition, we analyse the effect of the quantity of training data on the accuracy of the predictions and explore the effect of data augmentation using synthesized data.

****

Discriminative Deep Metric Learning for Face Verification in the Wild

Junlin Hu, Jiwen Lu, Yap-Peng Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1875-1882

This paper presents a new discriminative deep metric learning (DDML) method for face verification in the wild. Different from existing metric learning-based face verification methods which aim to learn a Mahalanobis distance metric to maximize the inter-class variations and minimize the intra-class variations, simultaneously, the proposed DDML trains a deep neural network which learns a set of hierarchical nonlinear transformations to project face pairs into the same feature subspace, under which the distance of each positive face pair is less than a smaller threshold and that of each negative pair is higher than a larger threshold, respectively, so that discriminative information can be exploited in the deep network. Our method achieves very competitive face verification performance on the widely used LFW and YouTube Faces (YTF) datasets.

****

Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses

Meina Kan, Shiguang Shan, Hong Chang, Xilin Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1883-1890

Identifying subjects with variations caused by poses is one of the most challenging tasks in face recognition, since the difference in appearances caused by poses may be even larger than the difference due to identity. Inspired by the observation that pose variations change non-linearly but smoothly, we propose to learn pose-robust features by modeling the complex non-linear transform from the non-frontal face images to frontal ones through a deep network in a progressive way, termed as stacked progressive auto-encoders (SPAE). Specifically, each shallow progressive auto-encoder of the stacked network is designed to map the face images at large poses to a virtual view at smaller ones, and meanwhile keep those images already at smaller poses unchanged. Then, stacking multiple these shallow auto-encoders can convert non-frontal face images to frontal ones progressively, which means the pose variations are narrowed down to zero step by step. As a result, the outputs of the topmost hidden layers of the stacked network contain very small pose variations, which can be used as the pose-robust features for face recognition. An additional attractiveness of the proposed method is that no pose estimation is needed for the test images. The proposed method is evaluated on two datasets with pose variations, i.e., MultiPIE and FERET datasets, and the experimental results demonstrate the superiority of our method to the existing works, especially to those 2D ones.

****

Deep Learning Face Representation from Predicting 10,000 Classes

Yi Sun, Xiaogang Wang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1891-1898

This paper proposes to learn a set of high-level feature representations through deep learning, referred to as Deep hidden IDentity features (DeepID), for face verification. We argue that DeepID can be effectively learned through challenging multi-class face identification tasks, whilst they can be generalized to other tasks (such as verification) and new identities unseen in the training set. Moreover, the generalization capability of DeepID increases as more face classes are to be predicted at training. DeepID features are taken from the last hidden layer neuron activations of deep convolutional networks (ConvNets). When learned as classifiers to recognize about 10,000 face identities in the training set and configured to keep reducing the neuron numbers along the feature extraction hierarchy, these deep ConvNets gradually form compact identity-related features in the top layers with only a small number of hidden neurons. The proposed features are extracted from various face regions to form complementary and over-complete representations. Any state-of-the-art classifiers can be learned based on these high-level representations for face verification. 97.45% verification accuracy on LFW is achieved with only weakly aligned faces.

****

3D-aided Face Recognition Robust to Expression and Pose Variations

Baptiste Chu, Sami Romdhani, Liming Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1899-1906

Expression and pose variations are major challenges for reliable face recognition (FR) in 2D. In this paper, we aim to endow state of the art face recognition SDKs with robustness to facial expression variations and pose changes by using an extended 3D Morphable Model (3DMM) which isolates identity variations from those due to facial expressions. Specifically, given a probe with expression, a novel view of the face is generated where the pose is rectified and the expression neutralized. We present two methods of expression neutralization. The first one uses prior knowledge to infer the neutral expression image from an input image. The second method, specifically designed for verification, is based on the transfer of the gallery face expression to the probe. Experiments using rectified and neutralized view with a standard commercial FR SDK on two 2D face databases, namely Multi-PIE and AR, show significant performance improvement of the commercial SDK to deal with expression and pose variations and demonstrates the effectiveness of the proposed approach.

****

Learning Non-Linear Reconstruction Models for Image Set Classification

Munawar Hayat, Mohammed Bennamoun, Senjian An; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1907-1914

We propose a deep learning framework for image set classification with application to face recognition. An Adaptive Deep Network Template (ADNT) is defined whose parameters are initialized by performing unsupervised pre-training in a layer-wise fashion using Gaussian Restricted Boltzmann Machines (GRBMs). The pre-initialized ADNT is then separately trained for images of each class and class-specific models are learnt. Based on the minimum reconstruction error from the learnt class-specific models, a majority voting strategy is used for classification. The proposed framework is extensively evaluated for the task of image set classification based face recognition on Honda/UCSD, CMU Mobo, YouTube Celebrities and a Kinect dataset. Our experimental results and comparisons with existing state-of-the-art methods show that the proposed method consistently achieves the best performance on all these datasets.

****

Gesture Recognition Portfolios for Personalization

Angela Yao, Luc Van Gool, Pushmeet Kohli; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1915-1922

Human gestures, similar to speech and handwriting, are often unique to the individual. Training a generic classifier applicable to everyone can be very difficult and as such, it has become a standard to use personalized classifiers in speech and handwriting recognition. In this paper, we address the problem of personalization in the context of gesture recognition, and propose a novel and extremely efficient way of doing personalization. Unlike conventional personalization methods which learn a single classifier that later gets adapted, our approach learns a set (portfolio) of classifiers during training, one of which is selected for each test subject based on the personalization data. We formulate classifier personalization as a selection problem and propose several algorithms to compute the set of candidate classifiers. Our experiments show that such an approach is much more efficient than adapting the classifier parameters but can still achieve comparable or better results.

****

Sign Spotting using Hierarchical Sequential Patterns with Temporal Intervals

Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, Richard Bowden; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1923-1930

This paper tackles the problem of spotting a set of signs occuring in videos with sequences of signs. To achieve this, we propose to model the spatio-temporal signatures of a sign using an extension of sequential patterns that contain temporal intervals called {\em Sequential Interval Patterns} (SIP). We then propose a novel multi-class classifier that organises different sequential interval patterns in a hierarchical tree structure called a Hierarchical SIP Tree (HSP-Tree). This allows one to exploit any subsequence sharing that exists between different SIPs of different classes. Multiple trees are then combined together into a forest of HSP-Trees resulting in a strong classifier that can be used to spot signs. We then show how the HSP-Forest can be used to spot sequences of signs that occur in an input video. We have evaluated the method on both concatenated sequences of isolated signs and continuous sign sequences. We also show that the proposed method is superior in robustness and accuracy to a state of the art sign recogniser when applied to spotting a sequence of signs.

****

Automatic Feature Learning for Robust Shadow Detection

Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, Roberto Togneri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1931-1938

We present a practical framework to automatically detect shadows in real world scenes from a single photograph. Previous works on shadow detection put a lot of effort in designing shadow variant and invariant hand-crafted features. In contrast, our framework automatically learns the most relevant features in a supervis

ed manner using multiple convolutional deep neural networks (ConvNets). The 7-layer network architecture of each ConvNet consists of alternating convolution and sub-sampling layers. The proposed framework learns features at the super-pixel level and along the object boundaries. In both cases, features are extracted usiing a context aware window centered at interest points. The predicted posteriors based on the learned features are fed to a conditional random field model to generate smooth shadow contours. Our proposed framework consistently performed better than the state-of-the-art on all major shadow databases collected under a variety of conditions.
****

Packing and Padding: Coupled Multi-index for Accurate Image Retrieval
Liang Zheng, Shengjin Wang, Ziqiong Liu, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1939-1946
In Bag-of-Words (BoW) based image retrieval, the SIFT visual word has a low discriminative power, so false positive matches occur prevalently. Apart from the information loss during quantization, another cause is that the SIFT feature only describes the local gradient distribution. To address this problem, this paper proposes a coupled Multi-Index (c-MI) framework to perform feature fusion at indexing level. Basically, complementary features are coupled into a multi-dimensional inverted index. Each dimension of c-MI corresponds to one kind of feature, and the retrieval process votes for images similar in both SIFT and other feature spaces. Specifically, we exploit the fusion of local color feature into c-MI. While the precision of visual match is greatly enhanced, we adopt Multiple Assignment to improve recall. The joint cooperation of SIFT and color features significantly reduces the impact of false positive matches.   Extensive experiments on several benchmark datasets demonstrate that c-MI improves the retrieval accuracy significantly, while consuming only half of the query time compared to the baseline. Importantly, we show that c-MI is well complementary to many prior techniques. Assembling these methods, we have obtained an mAP of 85.8% and N-S score of 3.85 on Holidays and Ukbench datasets, respectively, which compare favorably with the state-of-the-arts.
****

Adaptive Object Retrieval with Kernel Reconstructive Hashing
Haichuan Yang, Xiao Bai, Jun Zhou, Peng Ren, Zhihong Zhang, Jian Cheng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1947-1954
Hashing is very useful for fast approximate similarity search on large database. In the unsupervised settings, most hashing methods aim at preserving the similarity defined by Euclidean distance. Hash codes generated by these approaches only keep their Hamming distance corresponding to the pairwise Euclidean distance, ignoring the local distribution of each data point. This objective does not hold for k-nearest neighbors search. In this paper, we firstly propose a new adaptive similarity measure which is consistent with k-NN search, and prove that it leads to a valid kernel. Then we propose a hashing scheme which uses binary codes to preserve the kernel function. Using low-rank approximation, our hashing framework is more effective than existing methods that preserve similarity over arbitrary kernel. The proposed kernel function, hashing framework, and their combination have demonstrated significant advantages compared with several state-of-the-art methods.
****

Bayes Merging of Multiple Vocabularies for Scalable Image Retrieval
Liang Zheng, Shengjin Wang, Wengang Zhou, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1955-1962
In the Bag-of-Words (BoW) model, the vocabulary is of key importance. Typically, multiple vocabularies are generated to correct quantization artifacts and improve recall. However, this routine is corrupted by vocabulary correlation, i.e., overlapping among different vocabularies. Vocabulary correlation leads to an over-counting of the indexed features in the overlapped area, or the intersection set, thus compromising the retrieval accuracy. In order to address the correlation problem while preserve the benefit of high recall, this paper proposes a Bayes

merging approach to down-weight the indexed features in the intersection set. Through explicitly modeling the correlation problem in a probabilistic view, a joint similarity on both image- and feature-level is estimated for the indexed features in the intersection set. We evaluate our method on three benchmark datasets. Albeit simple, Bayes merging can be well applied in various merging tasks, and consistently improves the baselines on multi-vocabulary merging. Moreover, Bayes merging is efficient in terms of both time and memory cost, and yields competitive performance with the state-of-the-art methods.

****

Fast Supervised Hashing with Decision Trees for High-Dimensional Data

Guosheng Lin, Chunhua Shen, Qinfeng Shi, Anton van den Hengel, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1963-1970

Supervised hashing aims to map the original features to compact binary codes that are able to preserve label based similarity in the Hamming space. Non-linear hash functions have demonstrated their advantage over linear ones due to their powerful generalization capability. In the literature, kernel functions are typically used to achieve non-linearity in hashing, which achieve encouraging retrieval performance at the price of slow evaluation and training time. Here we propose to use boosted decision trees for achieving non-linearity in hashing, which are fast to train and evaluate, hence more suitable for hashing with high dimensional data. In our approach, we first propose sub-modular formulations for the hashing binary code inference problem and an efficient GraphCut based block search method for solving large-scale inference. Then we learn hash functions by training boosted decision trees to fit the binary codes. Experiments demonstrate that our proposed method significantly outperforms most state-of-the-art methods in retrieval precision and training time. Especially for high-dimensional data, our method is orders of magnitude faster than many methods in terms of training time.

****

Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, Alan Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1971-1978

Detecting objects becomes difficult when we need to deal with large shape deformation, occlusion and low resolution. We propose a novel approach to i) handle large deformations and partial occlusions in animals (as examples of highly deformable objects), ii) describe them in terms of body parts, and iii) detect them when their body parts are hard to detect (e.g., animals depicted at low resolution). We represent the holistic object and body parts separately and use a fully connected model to arrange templates for the holistic object and body parts. Our model automatically decouples the holistic object or body parts from the model when they are hard to detect. This enables us to represent a large number of holistic object and body part combinations to better deal with different "detectability" patterns caused by deformations, occlusion and/or low resolution. We apply our method to the six animal categories in the PASCAL VOC dataset and show that our method significantly improves state-of-the-art (by 4.1% AP) and provides a richer representation for objects. During training we use annotations for body parts (e.g., head, torso, etc.), making use of a new dataset of fully annotated object parts for PASCAL VOC 2010, which provides a mask for each part.

****

Associative Embeddings for Large-scale Knowledge Transfer with Self-assessment

Alexander Vezhnevets, Vittorio Ferrari; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1979-1986

We propose a method for knowledge transfer between semantically related classes in ImageNet. By transferring knowledge from the images that have bounding-box annotations to the others, our method is capable of automatically populating ImageNet with many more bounding-boxes. The underlying assumption that objects from semantically related classes look alike is formalized in our novel Associative Embedding (AE) representation. AE recovers the latent low-dimensional space of app earance variations among image windows. The dimensions of AE space tend to correspond to aspects of window appearance (e.g. side view, close up, background). We model the overlap of a window with an object using Gaussian Processes (GP) regression, which spreads annotation smoothly through AE space. The probabilistic nature of GP allows our method to perform self-assessment, i.e. assigning a quality estimate to its own output. It enables trading off the amount of returned annotations for their quality. A large scale experiment on 219 classes and 0.5 million images demonstrates that our method outperforms state-of-the-art methods and baselines for object localization. Using self-assessment we can automatically return bounding-box annotations for 51% of all images with high localization accuracy (i.e. 71% average overlap with ground-truth).

****

Detecting Objects using Deformation Dictionaries

Bharath Hariharan, C. L. Zitnick, Piotr Dollar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1987-1994

Several popular and effective object detectors separately model intra-class variations arising from deformations and appearance changes. This reduces model complexity while enabling the detection of objects across changes in view- point, object pose, etc. The Deformable Part Model (DPM) is perhaps the most successful such model to date. A common assumption is that the exponential number of templates enabled by a DPM is critical to its success. In this paper, we show the counter-intuitive result that it is possible to achieve similar accuracy using a small dictionary of deformations. Each component in our model is represented by a single HOG template and a dictionary of flow fields that determine the deformations the template may undergo. While the number of candidate deformations is dramatically fewer than that for a DPM, the deformed templates tend to be plausible and interpretable. In addition, we discover that the set of deformation bases is actually transferable across object categories and that learning shared bases across similar categories can boost accuracy.

****

Persistence-based Structural Recognition

Chunyuan Li, Maks Ovsjanikov, Frederic Chazal; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1995-2002

This paper presents a framework for object recognition using topological persistence. In particular, we show that the so-called persistence diagrams built from functions defined on the objects can serve as compact and informative descriptors for images and shapes. Complementary to the bag-of-features representation, which captures the distribution of values of a given function, persistence diagrams can be used to characterize its structural properties, reflecting spatial information in an invariant way. In practice, the choice of function is simple: each dimension of the feature vector can be viewed as a function. The proposed method is general: it can work on various multimedia data, including 2D shapes, textures and triangle meshes. Extensive experiments on 3D shape retrieval, hand gesture recognition and texture classification demonstrate the performance of the proposed method in comparison with state-of-the-art methods. Additionally, our approach yields higher recognition accuracy when used in conjunction with the bag-of-features.

****

Inferring Unseen Views of People

Chao-Yeh Chen, Kristen Grauman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2003-2010

We pose unseen view synthesis as a probabilistic tensor completion problem. Given images of people organized by their rough viewpoint, we form a 3D appearance tensor indexed by images (pose examples), viewpoints, and image positions. After discovering the low-dimensional latent factors that approximate that tensor, we can impute its missing entries. In this way, we generate novel synthetic views of peopleâ■■even when they are observed from just one camera viewpoint. We show that the inferred views are both visually and quantitatively accurate. Furthermore, we demonstrate their value for recognizing actions in unseen views and estimating viewpoint in novel images. While existing methods are often forced to choos

e between data that is either realistic or multi-view, our virtual views offer both, thereby allowing greater robustness to viewpoint in novel images.
****

Birdsnap: Large-scale Fine-grained Visual Categorization of Birds
Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, Peter N. Belhumeur; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2011-2018
We address the problem of large-scale fine-grained visual categorization, describing new methods we have used to produce an online field guide to 500 North American bird species. We focus on the challenges raised when such a system is asked to distinguish between highly similar species of birds. First, we introduce "one-vs-most classifiers." By eliminating highly similar species during training, these classifiers achieve more accurate and intuitive results than common one-vs-all classifiers. Second, we show how to estimate spatio-temporal class priors from observations that are sampled at irregular and biased locations. We show how these priors can be used to significantly improve performance. We then show state-of-the-art recognition performance on a new, large dataset that we make publicly available. These recognition methods are integrated into the online field guide, which is also publicly available.
****

Predicting Object Dynamics in Scenes
David F. Fouhey, C. L. Zitnick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2019-2026
Given a static scene, a human can trivially enumerate the myriad of things that can happen next and characterize the relative likelihood of each. In the process, we make use of enormous amounts of commonsense knowledge about how the world works. In this paper, we investigate learning this commonsense knowledge from data. To overcome a lack of densely annotated spatiotemporal data, we learn from sequences of abstract images gathered using crowdsourcing. The abstract scenes provide both object location and attribute information. We demonstrate qualitatively and quantitatively that our models produce plausible scene predictions on both the abstract images, as well as natural images taken from the Internet.
****

Enriching Visual Knowledge Bases via Object Discovery and Segmentation
Xinlei Chen, Abhinav Shrivastava, Abhinav Gupta; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2027-2034
There have been some recent efforts to build visual knowledge bases from Internet images. But most of these approaches have focused on bounding box representation of objects. In this paper, we propose to enrich these knowledge bases by automatically discovering objects and their segmentations from noisy Internet images. Specifically, our approach combines the power of generative modeling for segmentation with the effectiveness of discriminative models for detection. The key idea behind our approach is to learn and exploit top-down segmentation priors based on visual subcategories. The strong priors learned from these visual subcategories are then combined with discriminatively trained detectors and bottom up cues to produce clean object segmentations. Our experimental results indicate state-of-the-art performance on the difficult dataset introduced by Rubinstein et al. We have integrated our algorithm in NEIL for enriching its knowledge base. As of 14th April 2014, NEIL has automatically generated approximately 500K segmentations using web data.
****

Seeing the Arrow of Time
Lyndsey C. Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, William T. Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2035-2042
We explore whether we can observe Time's Arrow in a temporal sequence--is it possible to tell whether a video is running forwards or backwards? We investigate this somewhat philosophical question using computer vision and machine learning techniques. We explore three methods by which we might detect Time's Arrow in video sequences, based on distinct ways in which motion in video sequences might b

e asymmetric in time. We demonstrate good video forwards/backwards classification results on a selection of YouTube video clips, and on natively-captured sequences (with no temporally-dependent video compression), and examine what motions the models have learned that help discriminate forwards from backwards time.

****

Hierarchical Feature Hashing for Fast Dimensionality Reduction

Bin Zhao, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2043-2050

Curse of dimensionality is a practical and challenging problem in image categorization, especially in cases with a large number of classes. Multi-class classification encounters severe computational and storage problems when dealing with these large scale tasks. In this paper, we propose hierarchical feature hashing to effectively reduce dimensionality of parameter space without sacrificing classification accuracy, and at the same time exploit information in semantic taxonomy among categories. We provide detailed theoretical analysis on our proposed hashing method. Moreover, experimental results on object recognition and scene classification further demonstrate the effectiveness of hierarchical feature hashing.

****

Modeling Image Patches with a Generic Dictionary of Mini-Epitomes

George Papandreou, Liang-Chieh Chen, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2051-2058

The goal of this paper is to question the necessity of features like SIFT in categorical visual recognition tasks. As an alternative, we develop a generative model for the raw intensity of image patches and show that it can support image classification performance on par with optimized SIFT-based techniques in a bag-of-visual-words setting. Key ingredient of the proposed model is a compact dictionary of mini-epitomes, learned in an unsupervised fashion on a large collection of images. The use of epitomes allows us to explicitly account for photometric and position variability in image appearance. We show that this flexibility considerably increases the capacity of the dictionary to accurately approximate the appearance of image patches and support recognition tasks. For image classification, we develop histogram-based image encoding methods tailored to the epitomic representation, as well as an "epitomic footprint" encoding which is easy to visualize and highlights the generative nature of our model. We discuss in detail computational aspects and develop efficient algorithms to make the model scalable to large tasks. The proposed techniques are evaluated with   experiments on the challenging PASCAL VOC 2007 image classification benchmark.

****

Simplex-Based 3D Spatio-Temporal Feature Description for Action Recognition

Hao Zhang, Wenjun Zhou, Christopher Reardon, Lynne E. Parker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2059-2066

We present a novel feature description algorithm to describe 3D local spatio-temporal features for human action recognition. Our descriptor avoids the singularity and limited discrimination power issues of traditional 3D descriptors by quantizing and describing visual features in the simplex topological vector space. Specifically, given a feature's support region containing a set of 3D visual cues, we decompose the cues' orientation into three angles, transform the decomposed angles into the simplex space, and describe them in such a space. Then, quadrant decomposition is performed to improve discrimination, and a final feature vector is composed from the resulting histograms. We develop intuitive visualization tools for analyzing feature characteristics in the simplex topological vector space. Experimental results demonstrate that our novel simplex-based orientation decomposition (SOD) descriptor substantially outperforms traditional 3D descriptors for the KTH, UCF Sport, and Hollywood-2 benchmark action datasets. In addition, the results show that our SOD descriptor is a superior individual descriptor for action recognition.

****

In Search of Inliers: 3D Correspondence by Local and Global Voting

Anders Glent Buch, Yang Yang, Norbert Kruger, Henrik Gordon Petersen; Proceeding s of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2067-2074

We present a method for finding correspondence between 3D models. From an initial set of feature correspondences, our method uses a fast voting scheme to separate the inliers from the outliers. The novelty of our method lies in the use of a combination of local and global constraints to determine if a vote should be cast. On a local scale, we use simple, low-level geometric invariants. On a global scale, we apply covariant constraints for finding compatible correspondences. We guide the sampling for collecting voters by downward dependencies on previous voting stages. All of this together results in an accurate matching procedure. We evaluate our algorithm by controlled and comparative testing on different datasets, giving superior performance compared to state of the art methods. In a final experiment, we apply our method for 3D object detection, showing potential use of our method within higher-level vision.

****

Collective Matrix Factorization Hashing for Multimodal Data

Guiguang Ding, Yuchen Guo, Jile Zhou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2075-2082

Nearest neighbor search methods based on hashing have attracted considerable attention for effective and efficient large-scale similarity search in computer vision and information retrieval community. In this paper, we study the problems of learning hash functions in the context of multimodal data for cross-view similarity search. We put forward a novel hashing method, which is referred to Collective Matrix Factorization Hashing (CMFH). CMFH learns unified hash codes by collective matrix factorization with latent factor model from different modalities of one instance, which can not only supports cross-view search but also increases the search accuracy by merging multiple view information sources. We also prove that CMFH, a similarity-preserving hashing learning method, has upper and lower boundaries. Extensive experiments verify that CMFH significantly outperforms several state-of-the-art methods on three different datasets.

****

Finding Matches in a Haystack: A Max-Pooling Strategy for Graph Matching in the Presence of Outliers

Minsu Cho, Jian Sun, Olivier Duchenne, Jean Ponce; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2083-2090

A major challenge in real-world feature matching problems is to tolerate the numerous outliers arising in typical visual tasks. Variations in object appearance, shape, and structure within the same object class make it harder to distinguish inliers from outliers due to clutters. In this paper, we propose a max-pooling approach to graph matching, which is not only resilient to deformations but also remarkably tolerant to outliers. The proposed algorithm evaluates each candidate match using its most promising neighbors, and gradually propagates the corresponding scores to update the neighbors. As final output, it assigns a reliable score to each match together with its supporting neighbors, thus providing contextual information for further verification. We demonstrate the robustness and utility of our method with synthetic and real image experiments.

****

Locality in Generic Instance Search from One Example

Ran Tao, Efstratios Gavves, Cees G.M. Snoek, Arnold W.M. Smeulders; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2091-2098

This paper aims for generic instance search from a single example. Where the state-of-the-art relies on global image representation for the search, we proceed by including locality at all steps of the method. As the first novelty, we consider many boxes per database image as candidate targets to search locally in the picture using an efficient point-indexed representation. The same representation allows, as the second novelty, the application of very large vocabularies in the powerful Fisher vector and VLAD to search locally in the feature space. As the third novelty we propose an exponential similarity function to further emphasize locality in the feature space. Locality is advantageous in instance search as i t will rest on the matching unique details. We demonstrate a substantial increase in generic instance search performance from one example on three standard datasets with buildings, logos, and scenes from 0.443 to 0.620 in mAP.

****

Congruency-Based Reranking

Itai Ben-Shalom, Noga Levy, Lior Wolf, Nachum Dershowitz, Adiel Ben-Shalom, Roni Shweka, Yaacov Choueka, Tamir Hazan, Yaniv Bar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2099-2106

We present a tool for re-ranking the results of a specific query by considering the (n+1) × (n+1) matrix of pairwise similarities among the elements of the set of n retrieved results and the query itself. The re-ranking thus makes use of the similarities between the various results and does not employ additional sources of information. The tool is based on graphical Bayesian models, which reinforce retrieved items strongly linked to other retrievals, and on repeated clustering to measure the stability of the obtained associations. The utility of the tool is demonstrated within the context of visual search of documents from the Cairo Genizah and for retrieval of paintings by the same artist and in the same style.

****

Asymmetric Sparse Kernel Approximations for Large-scale Visual Search

Damek Davis, Jonathan Balzer, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2107-2114

We introduce an asymmetric sparse approximate embedding optimized for fast kernel comparison operations arising in large-scale visual search. In contrast to other methods that perform an explicit approximate embedding using kernel PCA followed by a distance compression technique in R^d, which loses information at both steps, our method utilizes the implicit kernel representation directly. In addition, we empirically demonstrate that our method needs no explicit training step and can operate with a dictionary of random exemplars from the dataset. We evaluate our method on three benchmark image retrieval datasets: SIFT1M, ImageNet, and 80M-TinyImages.

****

Locally Linear Hashing for Extracting Non-Linear Manifolds

Go Irie, Zhenguo Li, Xiao-Ming Wu, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2115-2122

Previous efforts in hashing intend to preserve data variance or pairwise affinity, but neither is adequate in capturing the manifold structures hidden in most visual data. In this paper, we tackle this problem by reconstructing the locally linear structures of manifolds in the binary Hamming space, which can be learned by locality-sensitive sparse coding. We cast the problem as a joint minimization of reconstruction error and quantization loss, and show that, despite its NP-hardness, a local optimum can be obtained efficiently via alternative optimization. Our method distinguishes itself from existing methods in its remarkable ability to extract the nearest neighbors of the query from the same manifold, instead of from the ambient space. On extensive experiments on various image benchmarks, our results improve previous state-of-the-art by 28-74% typically, and 627% on the Yale face data.

****

Active Frame, Location, and Detector Selection for Automated and Manual Video Annotation

Vasiliy Karasev, Avinash Ravichandran, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2123-2130

We describe an information-driven active selection approach to determine which detectors to deploy at which location in which frame of a video to minimize semantic class label uncertainty at every pixel, with the smallest computational cost that ensures a given uncertainty bound. We show minimal performance reduction compared to a "paragon" algorithm running all detectors at all locations in all frames, at a small fraction of the computational cost. Our method can handle uncertainty in the labeling mechanism, so it can handle both "oracles" (manual annotation) or noisy detectors (automated annotation).

****

Distance Encoded Product Quantization

Jae-Pil Heo, Zhe Lin, Sung-Eui Yoon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2131-2138

Many binary code embedding techniques have been proposed for large-scale approximate nearest neighbor search in computer vision. Recently, product quantization that encodes the cluster index in each subspace has been shown to provide impressive accuracy for nearest neighbor search. In this paper, we explore a simple question: is it best to use all the bit budget for encoding a cluster index in each subspace? We have found that as data points are located farther away from the centers of their clusters, the error of estimated distances among those points becomes larger. To address this issue, we propose a novel encoding scheme that distributes the available bit budget to encoding both the cluster index and the quantized distance between a point and its cluster center. We also propose two different distance metrics tailored to our encoding scheme. We have tested our method against the-state-of-the-art techniques on several well-known benchmarks, and found that our method consistently improves the accuracy over other tested methods. This result is achieved mainly because our method accurately estimates distances between two data points with the new binary codes and distance metric.

****

Collaborative Hashing

Xianglong Liu, Junfeng He, Cheng Deng, Bo Lang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2139-2146

Hashing technique has become a promising approach for fast similarity search. Most of existing hashing research pursue the binary codes for the same type of entities by preserving their similarities. In practice, there are many scenarios involving nearest neighbor search on the data given in matrix form, where two different types of, yet naturally associated entities respectively correspond to its two dimensions or views. To fully explore the duality between the two views, we propose a collaborative hashing scheme for the data in matrix form to enable fast search in various applications such as image search using bag of words and recommendation using user-item ratings. By simultaneously preserving both the entity similarities in each view and the interrelationship between views, our collaborative hashing effectively learns the compact binary codes and the explicit hash functions for out-of-sample extension in an alternating optimization way. Extensive evaluations are conducted on three well-known datasets for search inside a single view and search across different views, demonstrating that our proposed method outperforms state-of-the-art baselines, with significant accuracy gains ranging from 7.67% to 45.87% relatively.

****

Scalable Object Detection using Deep Neural Networks

Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2147-2154

Deep convolutional neural networks have recently achieved state-of-the-art performance on a number of image recognition benchmarks, including the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2012). The winning model on the localization sub-task was a network that predicts a single bounding box and a confidence score for each object category in the image. Such a model captures the whole-image context around the objects but cannot handle multiple instances of the same object in the image without naively replicating the number of outputs for each instance. In this work, we propose a saliency-inspired neural network model for detection, which predicts a set of class-agnostic bounding boxes along with a single score for each box, corresponding to its likelihood of containing any object of interest. The model naturally handles a variable number of instances for each class and allows for cross-class generalization at the highest levels of the network. We are able to obtain competitive recognition performance on VOC2007 and ILSVRC2012, while using only the top few predicted locations in each image and a small number of neural network evaluations.

****

Multiview Shape and Reflectance from Natural Illumination

Geoffrey Oxholm, Ko Nishino; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2155-2162

The world is full of objects with complex reflectances, situated in complex illumination environments. Past work on full 3D geometry recovery, however, has tried to handle this complexity by framing it into simplistic models of reflectance (Lambetian, mirrored, or diffuse plus specular) or illumination (one or more point light sources). Though there has been some recent progress in directly utilizing such complexities for recovering a single view geometry, it is not clear how such single-view methods can be extended to reconstruct the full geometry. To this end, we derive a probabilistic geometry estimation method that fully exploits the rich signal embedded in complex appearance. Though each observation provides partial and unreliable information, we show how to estimate the reflectance responsible for the diverse appearance, and unite the orientation cues embedded in each observation to reconstruct the underlying geometry. We demonstrate the effectiveness of our method on synthetic and real-world objects. The results show that our method performs accurately across a wide range of real-world environments and reflectances that lies between the extremes that have been the focus of past work.

****

Reflectance and Fluorescent Spectra Recovery based on Fluorescent Chromaticity Invariance under Varying Illumination

Ying Fu, Antony Lam, Yasuyuki Kobashi, Imari Sato, Takahiro Okabe, Yoichi Sato; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2163-2170

In recent years, fluorescence analysis of scenes has received attention. Fluorescence can provide additional information about scenes, and has been used in applications such as camera spectral sensitivity estimation, 3D reconstruction, and color relighting. In particular, hyperspectral images of reflective-fluorescent scenes provide a rich amount of data. However, due to the complex nature of fluorescence, hyperspectral imaging methods rely on specialized equipment such as hyperspectral cameras and specialized illuminants. In this paper, we propose a more practical approach to hyperspectral imaging of reflective-fluorescent scenes using only a conventional RGB camera and varied colored illuminants. The key idea of our approach is to exploit a unique property of fluorescence: the chromaticity of fluorescence emissions are invariant under different illuminants. This allows us to robustly estimate spectral reflectance and fluorescence emission chromaticity. We then show that given the spectral reflectance and fluorescent chromaticity, the fluorescence absorption and emission spectra can also be estimated. We demonstrate in results that all scene spectra can be accurately estimated from RGB images. Finally, we show that our method can be used to accurately relight scenes under novel lighting.

****

What Camera Motion Reveals About Shape With Unknown BRDF

Manmohan Chandraker; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2171-2178

Psychophysical studies show motion cues inform about shape even with unknown reflectance. Recent works in computer vision have considered shape recovery for an object of unknown BRDF using light source or object motions. This paper addresses the remaining problem of determining shape from the (small or differential) motion of the camera, for unknown isotropic BRDFs. Our theory derives a differential stereo relation that relates camera motion to depth of a surface with unknown isotropic BRDF, which generalizes traditional Lambertian assumptions. Under orthographic projection, we show shape may not be constrained in general, but two motions suffice to yield an invariant for several restricted (still unknown) BRDFs exhibited by common materials. For the perspective case, we show that three differential motions suffice to yield surface depth for unknown isotropic BRDF and unknown directional lighting, while additional constraints are obtained with restrictions on BRDF or lighting. The limits imposed by our theory are intrinsic to the shape recovery problem and independent of choice of reconstruction method.

We outline with experiments how potential reconstruction methods may exploit our theory. We illustrate trends shared by theories on shape from motion of light, object or camera, relating reconstruction hardness to imaging complexity.
****

Photometric Stereo using Constrained Bivariate Regression for General Isotropic Surfaces
Satoshi Ikehata, Kiyoharu Aizawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2179-2186

This paper presents a photometric stereo method that is purely pixelwise and handles general isotropic surfaces in a stable manner. Following the recently proposed sum-of-lobes representation of the isotropic reflectance function, we constructed a constrained bivariate regression problem where the regression function is approximated by smooth, bivariate Bernstein polynomials. The unknown normal vector was separated from the unknown reflectance function by considering the inverse representation of the image formation process, and then we could accurately compute the unknown surface normals by solving a simple and efficient quadratic programming problem. Extensive evaluations that showed the state-of-the-art performance using both synthetic and real-world images were performed.
****

Robust Separation of Reflection from Multiple Images
Xiaojie Guo, Xiaochun Cao, Yi Ma; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2187-2194

When one records a video/image sequence through a transparent medium (e.g. glass), the image is often a superposition of a transmitted layer (scene behind the medium) and a reflected layer. Recovering the two layers from such images seems to be a highly ill-posed problem since the number of unknowns to recover is twice as many as the given measurements. In this paper, we propose a robust method to separate these two layers from multiple images, which exploits the correlation of the transmitted layer across multiple images, and the sparsity and independence of the gradient fields of the two layers. A novel Augmented Lagrangian Multiplier based algorithm is designed to efficiently and effectively solve the decomposition problem. The experimental results on both simulated and real data demonstrate the superior performance of the proposed method over the state of the arts, in terms of accuracy and simplicity.
****

Surface-from-Gradients: An Approach Based on Discrete Geometry Processing
Wuyuan Xie, Yunbo Zhang, Charlie C. L. Wang, Ronald C.-K. Chung; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2195-2202

In this paper, we propose an efficient method to reconstruct surface-from-gradients (SfG). Our method is formulated under the framework of discrete geometry processing. Unlike the existing SfG approaches, we transfer the continuous reconstruction problem into a discrete space and efficiently solve the problem via a sequence of least-square optimization steps. Our discrete formulation brings three advantages: 1) the reconstruction preserves sharp-features, 2) sparse/incomplete set of gradients can be well handled, and 3) domains of computation can have irregular boundaries. Our formulation is direct and easy to implement, and the comparisons with state-of-the-arts show the effectiveness of our method.
****

Socially-aware Large-scale Crowd Forecasting
Alexandre Alahi, Vignesh Ramanathan, Li Fei-Fei; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2203-2210

In crowded spaces such as city centers or train stations, human mobility looks complex, but is often influenced only by a few causes. We propose to quantitatively study crowded environments by introducing a dataset of 42 million trajectories collected in train stations. Given this dataset, we address the problem of forecasting pedestrians' destinations, a central problem in understanding large-scale crowd mobility. We need to overcome the challenges posed by a limited number of observations (e.g. sparse cameras), and change in pedestrian appearance cues across different cameras. In addition, we often have restrictions in the way pe destrians can move in a scene, encoded as priors over origin and destination (OD) preferences. We propose a new descriptor coined as Social Affinity Maps (SAM) to link broken or unobserved trajectories of individuals in the crowd, while using the OD-prior in our framework. Our experiments show improvement in performance through the use of SAM features and OD prior. To the best of our knowledge, our work is one of the first studies that provides encouraging results towards a better understanding of crowd behavior at the scale of million pedestrians.

****

L0 Regularized Stationary Time Estimation for Crowd Group Analysis

Shuai Yi, Xiaogang Wang, Cewu Lu, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2211-2218

We tackle stationary crowd analysis in this paper, which is similarly important as modeling mobile groups in crowd scenes and finds many applications in surveillance. Our key contribution is to propose a robust algorithm of estimating how long a foreground pixel becomes stationary. It is much more challenging than only subtracting background because failure at a single frame due to local movement of objects, lighting variation, and occlusion could lead to large errors on stationary time estimation. To accomplish decent results, sparse constraints along spatial and temporal dimensions are jointly added by mixed partials to shape a 3D stationary time map. It is formulated as a L0 optimization problem. Besides background subtraction, it distinguishes among different foreground objects, which are close or overlapped in the spatio-temporal space by using a locally shared foreground codebook. The proposed technologies are used to detect four types of stationary group activities and analyze crowd scene structures. We provide the first public benchmark dataset for stationary time estimation and stationary group analysis.

****

Scene-Independent Group Profiling in Crowd

Jing Shao, Chen Change Loy, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2219-2226

Groups are the primary entities that make up a crowd. Understanding group-level dynamics and properties is thus scientifically important and practically useful in a wide range of applications, especially for crowd understanding. In this study we show that fundamental group-level properties, such as intra-group stability and inter-group conflict, can be systematically quantified by visual descriptors. This is made possible through learning a novel Collective Transition prior, which leads to a robust approach for group segregation in public spaces. From the prior, we further devise a rich set of group property visual descriptors. These descriptors are scene-independent, and can be effectively applied to public-scene with variety of crowd densities and distributions. Extensive experiments on hundreds of public scene video clips demonstrate that such property descriptors are not only useful but also necessary for group state analysis and crowd scene understanding.

****

Temporal Sequence Modeling for Video Event Detection

Yu Cheng, Quanfu Fan, Sharath Pankanti, Alok Choudhary; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2227-2234

We present a novel approach for event detection in video by temporal sequence modeling. Exploiting temporal information has lain at the core of many approaches for video analysis (i.e., action, activity and event recognition). Unlike previous works doing temporal modeling at semantic event level, we propose to model temporal dependencies in the data at sub-event level without using event annotations. This frees our model from ground truth and addresses several limitations in previous work on temporal modeling. Based on this idea, we represent a video by a sequence of visual words learnt from the video, and apply the Sequence Memoizer [21] to capture long-range dependencies in a temporal context in the visual sequence. This data-driven temporal model is further integrated with event classification for jointly performing segmentation and classification of events in a video. We demonstrate the efficacy of our approach on two challenging datasets for

visual recognition.
****

Recognition of Complex Events: Exploiting Temporal Dynamics between Underlying Concepts

Subhabrata Bhattacharya, Mahdi M. Kalayeh, Rahul Sukthankar, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2235-2242

While approaches based on bags of features excel at low-level action classification, they are ill-suited for recognizing complex events in video, where concept-based temporal representations currently dominate. This paper proposes a novel representation that captures the temporal dynamics of windowed mid-level concept detectors in order to improve complex event recognition. We first express each video as an ordered vector time series, where each time step consists of the vector formed from the concatenated confidences of the pre-trained concept detectors. We hypothesize that the dynamics of time series for different instances from the same event class, as captured by simple linear dynamical system (LDS) models, are likely to be similar even if the instances differ in terms of low-level visual features. We propose a two-part representation composed of fusing: (1) a singular value decomposition of block Hankel matrices (SSID-S) and (2) a harmonic signature (HS) computed from the corresponding eigen-dynamics matrix. The proposed method offers several benefits over alternate approaches: our approach is straightforward to implement, directly employs existing concept detectors and can be plugged into linear classification frameworks. Results on standard datasets such as NIST's TRECVID Multimedia Event Detection task demonstrate the improved accuracy of the proposed method.
****

Video Event Detection by Inferring Temporal Instance Labels

Kuan-Ting Lai, Felix X. Yu, Ming-Syan Chen, Shih-Fu Chang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2243-2250

Video event detection allows intelligent indexing of video content based on events. Traditional approaches extract features from video frames or shots, then quantize and pool the features to form a single vector representation for the entire video. Though simple and efficient, the final pooling step may lead to loss of temporally local information, which is important in indicating which part in a long video signifies presence of the event. In this work, we propose a novel instance-based video event detection approach. We represent each video as multiple "instances", defined as video segments of different temporal intervals. The objective is to learn an instance-level event detection model based on only video-level labels. To solve this problem, we propose a large-margin formulation which treats the instance labels as hidden latent variables, and simultaneously infers the instance labels as well as the instance-level classification model. Our framework infers optimal solutions that assume positive videos have a large number of positive instances while negative videos have the fewest ones. Extensive experiments on large-scale video event datasets demonstrate significant performance gains. The proposed method is also useful in explaining the detection results by localizing the temporal segments in a video which is responsible for the positive detection.
****

Backscatter Compensated Photometric Stereo with 3 Sources

Chourmouzios Tsiotsios, Maria E. Angelopoulou, Tae-Kyun Kim, Andrew J. Davison; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2251-2258

Photometric stereo offers the possibility of object shape reconstruction via reasoning about the amount of light reflected from oriented surfaces. However, in murky media such as sea water, the illuminating light interacts with the medium and some of it is backscattered towards the camera. Due to this additive light component, the standard Photometric Stereo equations lead to poor quality shape estimation. Previous authors have attempted to reformulate the approach but have either neglected backscatter entirely or disregarded its non-uniformity on the se nsor when camera and lights are close to each other. We show that by compensating effectively for the backscatter component, a linear formulation of Photometric Stereo is allowed which recovers an accurate normal map using only 3 lights. Our backscatter compensation method for point-sources can be used for estimating the uneven backscatter directly from single images without any prior knowledge about the characteristics of the medium or the scene. We compare our method with previous approaches through extensive experimental results, where a variety of objects are imaged in a big water tank whose turbidity is systematically increased, and show reconstruction quality which degrades little relative to clean water results even with a very significant scattering level.

****

Calibrating a Non-isotropic Near Point Light Source using a Plane
Jaesik Park, Sudipta N. Sinha, Yasuyuki Matsushita, Yu-Wing Tai, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2259-2266

We show that a non-isotropic near point light source rigidly attached to a camera can be calibrated using multiple images of a weakly textured planar scene. We prove that if the radiant intensity distribution (RID) of a light source is radially symmetric with respect to its dominant direction, then the shading observed on a Lambertian scene plane is bilaterally symmetric with respect to a 2D line on the plane. The symmetry axis detected in an image provides a linear constraint for estimating the dominant light axis. The light position and RID parameters can then be estimated using a linear method. Specular highlights if available can also be used for light position estimation. We also extend our method to handle non-Lambertian reflectances which we model using a biquadratic BRDF. We have evaluated our method on synthetic data quantitavely. Our experiments on real scenes show that our method works well in practice and enables light calibration without the need of a specialized hardware.

****

A New Perspective on Material Classification and Ink Identification
Rakesh Shiradkar, Li Shen, George Landon, Sim Heng Ong, Ping Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2267-2274

The surface bi-directional reflectance distribution function (BRDF) can be used to distinguish different materials. The BRDFs of many real materials are near isotropic and can be approximated well by a 2D function. When the camera principal axis is coincident with the surface normal of the material sample, the captured BRDF slice is nearly 1D, which suffers from significant information loss. Thus, improvement in classification performance can be achieved by simply setting the camera at a slanted view to capture a larger portion of the BRDF domain. We further use a handheld flashlight camera to capture a 1D BRDF slice for material classification. This 1D slice captures important reflectance properties such as specular reflection and retro-reflectance. We apply these results on ink classification, which can be used in forensics and analyzing historical manuscripts. For the first time, we show that most of the inks on the market can be well distinguished by their reflectance properties.

****

High Quality Photometric Reconstruction using a Depth Camera
Sk. Mohammadul Haque, Avishek Chatterjee, Venu Madhav Govindu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2275-2282

In this paper we present a depth-guided photometric 3D reconstruction method that works solely with a depth camera like the Kinect. Existing methods that fuse depth with normal estimates use an external RGB camera to obtain photometric information and treat the depth camera as a black box that provides a low quality depth estimate. Our contribution to such methods are two fold. Firstly, instead of using an extra RGB camera, we use the infra-red (IR) camera of the depth camera system itself to directly obtain high resolution photometric information. We believe that ours is the first method to use an IR depth camera system in this manner. Secondly, photometric methods applied to complex objects result in numerous holes in the reconstructed surface due to shadows and self-occlusions. To mitigate this problem, we develop a simple and effective multiview reconstruction approach that fuses depth and normal information from multiple viewpoints to build a complete, consistent and accurate 3D surface representation. We demonstrate the efficacy of our method to generate high quality 3D surface reconstructions for some complex 3D figurines.

****

Robust Surface Reconstruction via Triple Sparsity

Hicham Badri, Hussein Yahia, Driss Aboutajdine; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2283-2290

Reconstructing a surface/image from corrupted gradient fields is a crucial step in many imaging applications where a gradient field is subject to both noise and unlocalized outliers, resulting typically in a non-integrable field. We present in this paper a new optimization method for robust surface reconstruction. The proposed formulation is based on a triple sparsity prior : a sparse prior on the residual gradient field and a double sparse prior on the surface gradients. We develop an efficient alternate minimization strategy to solve the proposed optimization problem. The method is able to recover a good quality surface from severely corrupted gradients thanks to its ability to handle both noise and outliers. We demonstrate the performance of the proposed method on synthetic and real data. Experiments show that the proposed solution outperforms some existing methods in the three possible cases : noise only, outliers only and mixed noise/outliers.

****

Scattering Parameters and Surface Normals from Homogeneous Translucent Materials using Photometric Stereo

Bo Dong, Kathleen D. Moore, Weiyi Zhang, Pieter Peers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2291-2298

This paper proposes a novel photometric stereo solution to jointly estimate surface normals and scattering parameters from a globally planar, homogeneous, translucent object. Similar to classic photometric stereo, our method only requires as few as three observations of the translucent object under directional lighting. Naively applying classic photometric stereo results in blurred photometric normals. We develop a novel blind deconvolution algorithm based on inverse rendering for recovering the sharp surface normals and the material properties. We demonstrate our method on a variety of translucent objects.

****

Better Shading for Better Shape Recovery

Moumen T. El-Melegy, Aly S. Abdelrahim, Aly A. Farag; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2299-2304

The basic idea of shape from shading is to infer the shape of a surface from its shading information in a single image. Since this problem is ill-posed, a number of simplifying assumptions have been often used. However they rarely hold in practice. This paper presents a simple shading-correction algorithm that transforms the image to a new image that better satisfies the assumptions typically needed by existing algorithms, thus improving the accuracy of shape recovery. The algorithm takes advantage of some local shading measures that have been driven under these assumptions. The method is successfully evaluated on real data of human teeth with ground-truth 3D shapes.

****

Stable and Informative Spectral Signatures for Graph Matching

Nan Hu, Raif M. Rustamov, Leonidas Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2305-2312

In this paper, we consider the approximate weighted graph matching problem and introduce stable and informative first and second order compatibility terms suitable for inclusion into the popular integer quadratic program formulation. Our approach relies on a rigorous analysis of stability of spectral signatures based on the graph Laplacian. In the case of the first order term, we derive an objective function that measures both the stability and informativeness of a given spectral signature. By optimizing this objective, we design new spectral node signat ures tuned to a specific graph to be matched. We also introduce the pairwise heat kernel distance as a stable second order compatibility term; we justify its plausibility by showing that in a certain limiting case it converges to the classical adjacency matrix-based second order compatibility function. We have tested our approach on a set of synthetic graphs, the widely-used CMU house sequence, and a set of real images. These experiments show the superior performance of our first and second order compatibility terms as compared with the commonly used ones.

****

Deformable Object Matching via Deformation Decomposition based 2D Label MRF

Kangwei Liu, Junge Zhang, Kaiqi Huang, Tieniu Tan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2313-2320

Deformable object matching, which is also called elastic matching or deformation matching, is an important and challenging problem in computer vision. Although numerous deformation models have been proposed in different matching tasks, not many of them investigate the intrinsic physics underlying deformation. Due to the lack of physical analysis, these models cannot describe the structure changes of deformable objects very well. Motivated by this, we analyze the deformation physically and propose a novel deformation decomposition model to represent various deformations. Based on the physical model, we formulate the matching problem as a two-mensional label Markov Random Field. The MRF energy function is derived from the deformation decomposition model. Furthermore, we propose a two-stage method to optimize the MRF energy function. To provide a quantitative benchmark, we build a deformation matching database with an evaluation criterion. Experimental results show that our method outperforms previous approaches especially on complex deformations.

****

Locally Optimized Product Quantization for Approximate Nearest Neighbor Search

Yannis Kalantidis, Yannis Avrithis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2321-2328

We present a simple vector quantizer that combines low distortion with fast search and apply it to approximate nearest neighbor (ANN) search in high dimensional spaces. Leveraging the very same data structure that is used to provide non-exhaustive search, i.e., inverted lists or a multi-index, the idea is to locally optimize an individual product quantizer (PQ) per cell and use it to encode residuals. Local optimization is over rotation and space decomposition; interestingly, we apply a parametric solution that assumes a normal distribution and is extremely fast to train. With a reasonable space and time overhead that is constant in the data size, we set a new state-of-the-art on several public datasets, including a billion-scale one.

****

Multi-source Deep Learning for Human Pose Estimation

Wanli Ouyang, Xiao Chu, Xiaogang Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2329-2336

Visual appearance score, appearance mixture type and deformation are three important information sources for human pose estimation. This paper proposes to build a multi-source deep model in order to extract non-linear representation from these different aspects of information sources. With the deep model, the global, high-order human body articulation patterns in these information sources are extracted for pose estimation. The task for estimating body locations and the task for human detection are jointly learned using a unified deep model. The proposed approach can be viewed as a post-processing of pose estimation results and can flexibly integrate with existing methods by taking their information sources as input. By extracting the non-linear representation from multiple information sources, the deep model outperforms state-of-the-art by up to 8.6 percent on three public benchmark datasets.

****

Posebits for Monocular Human Pose Estimation

Gerard Pons-Moll, David J. Fleet, Bodo Rosenhahn; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2337-2344

We advocate the inference of qualitative information about 3D human pose, called posebits, from images. Posebits represent boolean geometric relationships between body parts (e.g., left-leg in front of right-leg or hands close to each other). The advantages of posebits as a mid-level representation are 1) for many tasks of interest, such qualitative pose information may be sufficient (e.g., semantic image retrieval), 2) it is relatively easy to annotate large image corpora with posebits, as it simply requires answers to yes/no questions; and 3) they help resolve challenging pose ambiguities and therefore facilitate the difficult talk of image-based 3D pose estimation. We introduce posebits, a posebit database, a method for selecting useful posebits for pose estimation and a structural SVM model for posebit inference. Experiments show the use of posebits for semantic image retrieval and for improving 3D pose estimation.
****
Real-time Simultaneous Pose and Shape Estimation for Articulated Objects Using a Single Depth Camera
Mao Ye, Ruigang Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2345-2352
In this paper we present a novel real-time algorithm for simultaneous pose and shape estimation for articulated objects, such as human beings and animals. The key of our pose estimation component is to embed the articulated deformation model with exponential-maps-based parametrization into a Gaussian Mixture Model. Benefiting from the probabilistic measurement model, our algorithm requires no explicit point correspondences as opposed to most existing methods. Consequently, our approach is less sensitive to local minimum and well handles fast and complex motions. Extensive evaluations on publicly available datasets demonstrate that our method outperforms most state-of-art pose estimation algorithms with large margin, especially in the case of challenging motions. Moreover, our novel shape adaptation algorithm based on the same probabilistic model automatically captures the shape of the subjects during the dynamic pose estimation process. Experiments show that our shape estimation method achieves comparable accuracy with state of the arts, yet requires neither parametric model nor extra calibration procedure.
****
Mixing Body-Part Sequences for Human Pose Estimation
Anoop Cherian, Julien Mairal, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2353-2360
In this paper, we present a method for estimating articulated human poses in videos. We cast this as an optimization problem defined on body parts with spatio-temporal links between them. The resulting formulation is unfortunately intractable and previous approaches only provide approximate solutions. Although such methods perform well on certain body parts, e.g., head, their performance on lower arms, i.e., elbows and wrists, remains poor. We present a new approximate scheme with two steps dedicated to pose estimation. First, our approach takes into account temporal links with subsequent frames for the less-certain parts, namely elbows and wrists. Second, our method decomposes poses into limbs, generates limb sequences across time, and recomposes poses by mixing these body part sequences . We introduce a new dataset "Poses in the Wild", which is more challenging than the existing ones, with sequences containing background clutter, occlusions, and severe camera motion. We experimentally compare our method with recent approaches on this new dataset as well as on two other benchmark datasets, and show significant improvement.
****
Robust Estimation of 3D Human Poses from a Single Image
Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L. Yuille, Wen Gao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2361-2368
Human pose estimation is a key step to action recognition. We propose a method of estimating 3D human poses from a single image, which works in conjunction with an existing 2D pose/joint detector. 3D pose estimation is challenging because m

ultiple 3D poses may correspond to the same 2D pose after projection due to the lack of depth information. Moreover, current 2D pose estimators are usually inaccurate which may cause errors in the 3D estimation. We address the challenges in three ways: (i) We represent a 3D pose as a linear combination of a sparse set of bases learned from 3D human skeletons. (ii) We enforce limb length constraints to eliminate anthropomorphically implausible skeletons. (iii) We estimate a 3D pose by minimizing the L_1-norm error between the projection of the 3D pose and the corresponding 2D detection. The L_1-norm loss term is robust to inaccurate 2D joint estimations. We use the alternating direction method (ADM) to solve the optimization problem efficiently. Our approach outperforms the state-of-the-arts on three benchmark datasets.

****

Fisher and VLAD with FLAIR

Koen E. A. van de Sande, Cees G. M. Snoek, Arnold W. M. Smeulders; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2369-2376

A major computational bottleneck in many current algorithms is the evaluation of arbitrary boxes. Dense local analysis and powerful bag-of-word encodings, such as Fisher vectors and VLAD, lead to improved accuracy at the expense of increased computation time. Where a simplification in the representation is tempting, we exploit novel representations while maintaining accuracy. We start from state-of-the-art, fast selective search, but our method will apply to any initial box-partitioning. By representing the picture as sparse integral images, one per codeword, we achieve a Fast Local Area Independent Representation. FLAIR allows for very fast evaluation of any box encoding and still enables spatial pooling. In FLAIR we achieve exact VLAD's difference coding, even with L2 and power-norms. Finally, by multiple codeword assignments, we achieve exact and approximate Fisher vectors with FLAIR. The results are a 18x speedup, which enables us to set a new state-of-the-art on the challenging 2010 PASCAL VOC objects and the fine-grained categorization of the CUB-2011 200 bird species. Plus, we rank number one in the official ImageNet 2013 detection challenge.

****

Immediate, Scalable Object Category Detection

Yusuf Aytar, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2377-2384

The objective of this work is object category detection in large scale image datasets in the manner of Video Google â■■ an object category is specified by a HOG classifier template, and retrieval is immediate at run time. We make the following three contributions: (i) a new image representation based on mid-level discriminative patches, that is designed to be suited to immediate object category detection and inverted file indexing; (ii) a sparse representation of a HOG classifier using a set of mid-level discriminative classifier patches; and (iii) a fast method for spatial reranking images on their detections. We evaluate the detection method on the standard PASCAL VOC 2007 dataset, together with a 100K image subset of ImageNet, and demonstrate near state of the art detection performance at low ranks whilst maintaining immediate retrieval speeds. Applications are also demonstrated using an exemplar-SVM for pose matched retrieval.

****

Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model

Golnaz Ghiasi, Charless C. Fowlkes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2385-2392

The presence of occluders significantly impacts performance of systems for object recognition. However, occlusion is typically treated as an unstructured source of noise and explicit models for occluders have lagged behind those for object appearance and shape. In this paper we describe a hierarchical deformable part model for face detection and keypoint localization that explicitly models occlusions of parts. The proposed model structure makes it possible to augment positive training data with large numbers of synthetically occluded instances. This allows us to easily incorporate the statistics of occlusion patterns in a discri minatively trained model. We test the model on several benchmarks for keypoint localization including challenging sets featuring significant occlusion. We find that the addition of an explicit model of occlusion yields a system that outperforms existing approaches in keypoint localization accuracy.

****

Word Channel Based Multiscale Pedestrian Detection Without Image Resizing and Using Only One Classifier

Arthur Daniel Costea, Sergiu Nedevschi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2393-2400

Most pedestrian detection approaches that achieve high accuracy and precision rate and that can be used for real-time applications are based on histograms of gradient orientations. Usually multiscale detection is attained by resizing the image several times and by recomputing the image features or using multiple classifiers for different scales. In this paper we present a pedestrian detection approach that uses the same classifier for all pedestrian scales based on image features computed for a single scale. We go beyond the low level pixel-wise gradient orientation bins and use higher level visual words organized into Word Channels. Boosting is used to learn classification features from the integral Word Channels. The proposed approach is evaluated on multiple datasets and achieves outstanding results on the INRIA and Caltech-USA benchmarks. By using a GPU implementation we achieve a classification rate of over 10 million bounding boxes per second and a 16 FPS rate for multiscale detection in a 640Ã■480 image.

****

Parsing Occluded People

Golnaz Ghiasi, Yi Yang, Deva Ramanan, Charless C. Fowlkes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2401-2408

Occlusion poses a significant difficulty for object recognition due to the combinatorial diversity of possible occlusion patterns. We take a strongly supervised, non-parametric approach to modeling occlusion by learning deformable models with many local part mixture templates using large quantities of synthetically generated training data. This allows the model to learn the appearance of different occlusion patterns including figure-ground cues such as the shapes of occluding contours as well as the co-occurrence statistics of occlusion between neighboring parts. The underlying part mixture-structure also allows the model to capture coherence of object support masks between neighboring parts and make compelling predictions of figure-ground-occluder segmentations. We test the resulting model on human pose estimation under heavy occlusion and find it produces improved localization accuracy.

****

Multi-fold MIL Training for Weakly Supervised Object Localization

Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2409-2416

Object category localization is a challenging problem in computer vision. Standard supervised training requires bounding box annotations of object instances. This time-consuming annotation process is sidestepped in weakly supervised learning. In this case, the supervised information is restricted to binary labels that indicate the absence/presence of object instances in the image, without their locations. We follow a multiple-instance learning approach that iteratively trains the detector and infers the object locations in the positive training images. Our main contribution is a multi-fold multiple instance learning procedure, which prevents training from prematurely locking onto erroneous object locations. This procedure is particularly important when high-dimensional representations, such as the Fisher vectors, are used. We present a detailed experimental evaluation using the PASCAL VOC 2007 dataset. Compared to state-of-the-art weakly supervised detectors, our approach better localizes objects in the training images, which translates into improved detection performance.

****

Generating Object Segmentation Proposals using Global and Local Search

Pekka Rantalankila, Juho Kannala, Esa Rahtu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2417-2424
We present a method for generating object segmentation proposals from groups of superpixels. The goal is to propose accurate segmentations for all objects of an image. The proposed object hypotheses can be used as input to object detection systems and thereby improve efficiency by replacing exhaustive search. The segmentations are generated in a class-independent manner and therefore the computational cost of the approach is independent of the number of object classes. Our approach combines both global and local search in the space of sets of superpixels. The local search is implemented by greedily merging adjacent pairs of superpixels to build a bottom-up segmentation hierarchy. The regions from such a hierarchy directly provide a part of our region proposals. The global search provides the other part by performing a set of graph cut segmentations on a superpixel graph obtained from an intermediate level of the hierarchy. The parameters of the graph cut problems are learnt in such a manner that they provide complementary sets of regions. Experiments with Pascal VOC images show that we reach state-of-the-art with greatly reduced computational cost.
****

A Novel Chamfer Template Matching Method Using Variational Mean Field
Duc Thanh Nguyen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2425-2432
This paper proposes a novel mean field-based Chamfer template matching method. In our method, each template is represented as a field model and matching a template with an input image is formulated as estimation of a maximum of posteriori in the field model. Variational approach is then adopted to approximate the estimation. The proposed method was applied for two different variants of Chamfer template matching and evaluated through the task of object detection. Experimental results on benchmark datasets including ETHZShapeClass and INRIAHorse have shown that the proposed method could significantly improve the accuracy of template matching while not sacrificing much of the efficiency. Comparisons with other recent template matching algorithms have also shown the robustness of the proposed method.
****

Confidence-Rated Multiple Instance Boosting for Object Detection
Karim Ali, Kate Saenko; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2433-2440
Over the past years, Multiple Instance Learning (MIL) has proven to be an effective framework for learning with weakly labeled data. Applications of MIL to object detection, however, were limited to handling the uncertainties of manual annotations. In this paper, we propose a new MIL method for object detection that is capable of handling the noisier automatically obtained annotations. Our approach consists in first obtaining confidence estimates over the label space and, second, incorporating these estimates within a new Boosting procedure. We demonstrate the efficiency of our procedure on two detection tasks, namely, horse detection and pedestrian detection, where the training data is primarily annotated by a coarse area of interest detector. We show dramatic improvements over existing MIL methods. In both cases, we demonstrate that an efficient appearance model can be learned using our approach.
****

COSTA: Co-Occurrence Statistics for Zero-Shot Classification
Thomas Mensink, Efstratios Gavves, Cees G.M. Snoek; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2441-2448
In this paper we aim for zero-shot classification, that is visual recognition of an unseen class by using knowledge transfer from known classes. Our main contribution is COSTA, which exploits co-occurrences of visual concepts in images for knowledge transfer. These inter-dependencies arise naturally between concepts, and are easy to obtain from existing annotations or web-search hit counts. We estimate a classifier for a new label, as a weighted combination of related classes, using the co-occurrences to define the weight.■ We propose various metrics to leverage these co-occurrences, and a regression model for learning a weight for

each related class. We also show that our zero-shot classifiers can serve as priors for few-shot learning. Experiments on three multi-labeled datasets reveal that our proposed zero-shot methods, are approaching and occasionally outperforming fully supervised SVMs. We conclude that co-occurrence statistics suffice for zero-shot classification.
****

Analysis by Synthesis: 3D Object Recognition by Object Reconstruction
Mohsen Hejrati, Deva Ramanan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2449-2456

We introduce a new approach for recognizing and reconstructing 3D objects in images. Our approach is based on an analysis by synthesis strategy. A forward synthesis model constructs possible geometric interpretations of the world, and then selects the interpretation that best agrees with the measured visual evidence. The forward model synthesizes visual templates defined on invariant (HOG) features. These visual templates are discriminatively trained to be accurate for inverse estimation. We introduce an efficient "brute-force" approach to inference that searches through a large number of candidate reconstructions, returning the optimal one. One benefit of such an approach is that recognition is inherently (re)constructive. We show state of the art performance for detection and reconstruction on two challenging 3D object recognition datasets of cars and cuboids.
****

Submodular Object Recognition
Fan Zhu, Zhuolin Jiang, Ling Shao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2457-2464

We present a novel object recognition framework based on multiple figure-ground hypotheses with a large object spatial support, generated by bottom-up processes and mid-level cues in an unsupervised manner. We exploit the benefit of regression for discriminating segments' categories and qualities, where a regressor is trained to each category using the overlapping observations between each figure-ground segment hypothesis and the ground-truth of the target category in an image. Object recognition is achieved by maximizing a submodular objective function, which maximizes the similarities between the selected segments (i.e., facility locations) and their group elements (i.e., clients), penalizes the number of selected segments, and more importantly, encourages the consistency of object categories corresponding to maximum regression values from different category-specific regressors for the selected segments. The proposed framework achieves impressive recognition results on three benchmark datasets, including PASCAL VOC 2007, Caltech-101 and ETHZ-shape.
****

Multimodal Learning in Loosely-organized Web Images
Kun Duan, David J. Crandall, Dhruv Batra; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2465-2472

Photo-sharing websites have become very popular in the last few years, leading to huge collections of online images. In addition to image data, these websites collect a variety of multimodal metadata about photos including text tags, captions, GPS coordinates, camera metadata, user profiles, etc. However, this metadata is not well constrained and is often noisy, sparse, or missing altogether. In this paper, we propose a framework to model these "loosely organized" multimodal datasets, and show how to perform loosely-supervised learning using a novel latent Conditional Random Field framework. We learn parameters of the LCRF automatically from a small set of validation data, using Information Theoretic Metric Learning (ITML) to learn distance functions and a structural SVM formulation to learn the potential functions. We apply our framework on four datasets of images from Flickr, evaluating both qualitatively and quantitatively against several baselines.
****

Generalized Max Pooling
Naila Murray, Florent Perronnin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2473-2480

State-of-the-art patch-based image representations involve a pooling operation t hat aggregates statistics computed from local descriptors. Standard pooling operations include sum- and max-pooling. Sum-pooling lacks discriminability because the resulting representation is strongly influenced by frequent yet often uninformative descriptors, but only weakly influenced by rare yet potentially highly-informative ones. Max-pooling equalizes the influence of frequent and rare descriptors but is only applicable to representations that rely on count statistics, such as the bag-of-visual-words (BOV) and its soft- and sparse-coding extensions. We propose a novel pooling mechanism that achieves the same effect as max-pooling but is applicable beyond the BOV and especially to the state-of-the-art Fisher Vector-- hence the name Generalized Max Pooling (GMP). It involves equalizing the similarity between each patch and the pooled representation, which is shown to be equivalent to re-weighting the per-patch statistics. We show on five public image classification benchmarks that the proposed GMP can lead to significant performance gains with respect to heuristic alternatives.

****

Domain Adaptation on the Statistical Manifold

Mahsa Baktashmotlagh, Mehrtash T. Harandi, Brian C. Lovell, Mathieu Salzmann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2481-2488

In this paper, we tackle the problem of unsupervised domain adaptation for classification. In the unsupervised scenario where no labeled samples from the target domain are provided, a popular approach consists in transforming the data such that the source and target distributions become similar. To compare the two distributions, existing approaches make use of the Maximum Mean Discrepancy (MMD). However, this does not exploit the fact that probability distributions lie on a Riemannian manifold. Here, we propose to make better use of the structure of this manifold and rely on the distance on the manifold to compare the source and target distributions. In this framework, we introduce a sample selection method and a subspace-based method for unsupervised domain adaptation, and show that both these manifold-based techniques outperform the corresponding approaches based on the MMD. Furthermore, we show that our subspace-based approach yields state-of-the-art results on a standard object recognition benchmark.

****

Nonparametric Part Transfer for Fine-grained Recognition

Christoph Goring, Erik Rodner, Alexander Freytag, Joachim Denzler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2489-2496

In the following paper, we present an approach for fine-grained recognition based on a new part detection method. In particular, we propose a nonparametric label transfer technique which transfers part constellations from objects with similar global shapes. The possibility for transferring part annotations to unseen images allows for coping with a high degree of pose and view variations in scenarios where traditional detection models (such as deformable part models) fail. Our approach is especially valuable for fine-grained recognition scenarios where intraclass variations are extremely high, and precisely localized features need to be extracted. Furthermore, we show the importance of carefully designed visual extraction strategies, such as combination of complementary feature types and iterative image segmentation, and the resulting impact on the recognition performance. In experiments, our simple yet powerful approach achieves 35.9% and 57.8% accuracy on the CUB-2010 and 2011 bird datasets, which is the current best performance for these benchmarks.

****

The Fastest Deformable Part Model for Object Detection

Junjie Yan, Zhen Lei, Longyin Wen, Stan Z. Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2497-2504

This paper solves the speed bottleneck of deformable part model (DPM), while maintaining the accuracy in detection on challenging datasets. Three prohibitive steps in cascade version of DPM are accelerated, including 2D correlation between root filter and feature map, cascade part pruning and HOG feature extraction. For 2D correlation, the root filter is constrained to be low rank, so that 2D corr elation can be calculated by more efficient linear combination of 1D correlations. A proximal gradient algorithm is adopted to progressively learn the low rank filter in a discriminative manner. For cascade part pruning, neighborhood aware cascade is proposed to capture the dependence in neighborhood regions for aggressive pruning. Instead of explicit computation of part scores, hypotheses can be pruned by scores of neighborhoods under the first order approximation. For HOG feature extraction, look-up tables are constructed to replace expensive calculations of orientation partition and magnitude with simpler matrix index operations. Extensive experiments show that (a) the proposed method is 4 times faster than the current fastest DPM method with similar accuracy on Pascal VOC, (b) the proposed method achieves state-of-the-art accuracy on pedestrian and face detection task with frame-rate speed.

****

Unsupervised Learning of Dictionaries of Hierarchical Compositional Models

Jifeng Dai, Yi Hong, Wenze Hu, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2505-2512

This paper proposes an unsupervised method for learning dictionaries of hierarchical compositional models for representing natural images. Each model is in the form of a template that consists of a small group of part templates that are allowed to shift their locations and orientations relative to each other, and each part template is in turn a composition of Gabor wavelets that are also allowed to shift their locations and orientations relative to each other. Given a set of unannotated training images, a dictionary of such hierarchical templates are learned so that each training image can be represented by a small number of templates that are spatially translated, rotated and scaled versions of the templates in the learned dictionary. The learning algorithm iterates between the following two steps: (1) Image encoding by a template matching pursuit process that involves a bottom-up template matching sub-process and a top-down template localization sub-process. (2) Dictionary re-learning by a shared matching pursuit process. Experimental results show that the proposed approach is capable of learning meaningful templates, and the learned templates are useful for tasks such as domain adaption and image cosegmentation.

****

Quasi Real-Time Summarization for Consumer Videos

Bin Zhao, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2513-2520

With the widespread availability of video cameras, we are facing an ever-growing enormous collection of unedited and unstructured video data. Due to lack of an automatic way to generate summaries from this large collection of consumer videos, they can be tedious and time consuming to index or search. In this work, we propose online video highlighting, a principled way of generating short video summarizing the most important and interesting contents of an unedited and unstructured video, costly both time-wise and financially for manual processing. Specifically, our method learns a dictionary from given video using group sparse coding, and updates atoms in the dictionary on-the-fly. A summary video is then generated by combining segments that cannot be sparsely reconstructed using the learned dictionary. The online fashion of our proposed method enables it to process arbitrarily long videos and start generating summaries before seeing the end of the video. Moreover, the processing time required by our proposed method is close to the original video length, achieving quasi real-time summarization speed. Theoretical analysis, together with experimental results on more than 12 hours of surveillance and YouTube videos are provided, demonstrating the effectiveness of online video highlighting.

****

Gait Recognition under Speed Transition

Al Mansur, Yasushi Makihara, Rasyid Aqmar, Yasushi Yagi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2521-2528

This paper describes a method of gait recognition from image sequences wherein a subject is accelerating or decelerating. As a speed change occurs due to a change of pitch (the first-order derivative of a phase, namely, a gait stance) and/or stride, we model this speed change using a cylindrical manifold whose azimuth and height corresponds to the phase and the stride, respectively. A radial basis function (RBF) interpolation framework is used to learn subject specific mapping matrices for mapping from manifold to image space. Given an input image sequence of speed transited gait of a test subject, we estimate the mapping matrix of the test subject as well as the phase and stride sequence using an energy minimization framework considering the following three points: (1) fitness of the synthesized images to the input image sequence as well as to an eigenspace constructed by exemplars of training subjects; (2) smoothness of the phase and the stride sequence; and (3) pitch and stride fitness to the pitch-stride preference model. Using the estimated mapping matrix, we synthesize a constant-speed gait image sequence, and extract a conventional period-based gait feature from it for matching. We conducted experiments using real speed transited gait image sequences with 179 subjects and demonstrated the effectiveness of the proposed method.

****

Video Classification using Semantic Concept Co-occurrences

Shayan Modiri Assari, Amir Roshan Zamir, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2529-2536

We address the problem of classifying complex videos based on their content. A typical approach to this problem is performing the classification using semantic attributes, commonly termed concepts, which occur in the video. In this paper, we propose a contextual approach to video classification based on Generalized Maximum Clique Problem (GMCP) which uses the co-occurrence of concepts as the context model. To be more specific, we propose to represent a class based on the co-occurrence of its concepts and classify a video based on matching its semantic co-occurrence pattern to each class representation. We perform the matching using GMCP which finds the strongest clique of co-occurring concepts in a video. We argue that, in principal, the co-occurrence of concepts yields a richer representation of a video compared to most of the current approaches. Additionally, we propose a novel optimal solution to GMCP based on Mixed Binary Integer Programming (MBIP). The evaluations show our approach, which opens new opportunities for further research in this direction, outperforms several well established video classification methods.

****

Temporal Segmentation of Egocentric Videos

Yair Poleg, Chetan Arora, Shmuel Peleg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2537-2544

The use of wearable cameras makes it possible to record life logging egocentric videos. Browsing such long unstructured videos is time consuming and tedious. Segmentation into meaningful chapters is an important first step towards adding structure to egocentric videos, enabling efficient browsing, indexing and summarization of the long videos. Two sources of information for video segmentation are (i) the motion of the camera wearer, and (ii) the objects and activities recorded in the video. In this paper we address the motion cues for video segmentation. Motion based segmentation is especially difficult in egocentric videos when the camera is constantly moving due to natural head movement of the wearer. We propose a robust temporal segmentation of egocentric videos into a hierarchy of motion classes using a new Cumulative Displacement Curves. Unlike instantaneous motion vectors, segmentation using integrated motion vectors performs well even in dynamic and crowded scenes. No assumptions are made on the underlying scene structure and the method works in indoor as well as outdoor situations. We demonstrate the effectiveness of our approach using publicly available videos as well as choreographed videos. We also suggest an approach to detect the fixation of wearer's gaze in the walking portion of the egocentric videos.

****

Efficient Action Localization with Approximately Normalized Fisher Vectors

Dan Oneata, Jakob Verbeek, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2545-2552

The Fisher vector (FV) representation is a high-dimensional extension of the popular bag-of-word representation. Transformation of the FV by power and L2 normalizations has shown to significantly improve its performance, and led to state-of-the-art results for a range of image and video classification and retrieval tasks. These normalizations, however, render the representation non-additive over local descriptors. Combined with its high dimensionality, this makes the FV computationally expensive for the purpose of localization tasks. In this paper we present approximations to both these normalizations, which yield significant improvements in the memory and computational costs of the FV when used for localization. Second, we show how these approximations can be used to define upper-bounds on the score function that can be efficiently evaluated, which enables the use of branch-and-bound search as an alternative to exhaustive sliding window search. We present experimental evaluation results on classification and temporal localization of actions in videos. These show that the our approximations lead to a speedup of at least one order of magnitude, while maintaining state-of-the-art action recognition and localization performance.

****

Unsupervised Trajectory Modelling using Temporal Information via Minimal Paths

Brais Cancela, Alberto Iglesias, Marcos Ortega, Manuel G. Penedo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2553-2560

This paper presents a novel methodology for modelling pedestrian trajectories over a scene, based in the hypothesis that, when people try to reach a destination, they use the path that takes less time, taking into account environmental information like the type of terrain or what other people did before. Thus, a minimal path approach can be used to model human trajectory behaviour. We develop a modified Fast Marching Method that allows us to include both velocity and orientation in the Front Propagation Approach, without increasing its computational complexity. Combining all the information, we create a time surface that shows the time a target need to reach any given position in the scene. We also create different metrics in order to compare the time surface against the real behaviour. Experimental results over a public dataset prove the initial hypothesis' correctness.

****

A Hierarchical Context Model for Event Recognition in Surveillance Video

Xiaoyang Wang, Qiang Ji; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2561-2568

Due to great challenges such as tremendous intra-class variations and low image resolution, context information has been playing a more and more important role for accurate and robust event recognition in surveillance videos. The context information can generally be divided into the feature level context, the semantic level context, and the prior level context. These three levels of context provide crucial bottom-up, middle level, and top down information that can benefit the recognition task itself. Unlike existing researches that generally integrate the context information at one of the three levels, we propose a hierarchical context model that simultaneously exploits contexts at all three levels and systematically incorporate them into event recognition. To tackle the learning and inference challenges brought in by the model hierarchy, we develop complete learning and inference algorithms for the proposed hierarchical context model based on variational Bayes method. Experiments on VIRAT 1.0 and 2.0 Ground Datasets demonstrate the effectiveness of the proposed hierarchical context model for improving the event recognition performance even under great challenges like large intra-class variations and low image resolution.

****

DISCOVER: Discovering Important Segments for Classification of Video Events and Recounting

Chen Sun, Ram Nevatia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2569-2576

We propose a unified framework DISCOVER to simultaneously discover important segments, classify high-level events and generate recounting for large amounts of u nconstrained web videos. The motivation is our observation that many video events are characterized by certain important segments. Our goal is to find the important segments and capture their information for event classification and recounting. We introduce an evidence localization model where evidence locations are modeled as latent variables. We impose constraints on global video appearance, local evidence appearance and the temporal structure of the evidence. The model is learned via a max-margin framework and allows efficient inference. Our method does not require annotating sources of evidence, and is jointly optimized for event classification and recounting. Experimental results are shown on the challenging TRECVID 2013 MEDTest dataset.

****

Towards Good Practices for Action Video Encoding

Jianxin Wu, Yu Zhang, Weiyao Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2577-2584

High dimensional representations such as VLAD or FV have shown excellent accuracy in action recognition. This paper shows that a proper encoding built upon VLAD can achieve further accuracy boost with only negligible computational cost. We empirically evaluated various VLAD improvement technologies to determine good practices in VLAD-based video encoding. Furthermore, we propose an interpretation that VLAD is a maximum entropy linear feature learning process. Combining this new perspective with observed VLAD data distribution properties, we propose a simple, lightweight, but powerful bimodal encoding method. Evaluated on 3 benchmark action recognition datasets (UCF101, HMDB51 and Youtube), the bimodal encoding improves VLAD by large margins in action recognition.

****

Improving Semantic Concept Detection through the Dictionary of Visually-distinct Elements

Afshin Dehghan, Haroon Idrees, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2585-2592

A video captures a sequence and interactions of concepts that can be static, for instance, objects or scenes, or dynamic, such as actions. For large datasets containing hundreds of thousands of images or videos, it is impractical to manually annotate all the concepts, or all the instances of a single concept. However, a dictionary with visuallydistinct elements can be created automatically from unlabeled videos which can capture and express the entire dataset. The downside to this machine-discovered dictionary is meaninglessness, i.e., its elements are devoid of semantics and interpretation. In this paper, we present an approach that leverages the strengths of semantic concepts and the machine-discovered DOVE by learning a relationship between them. Since instances of a semantic concept share visual similarity, the proposed approach uses softconsensus regularization to learn the mapping that enforces instances from each semantic concept to have similar representations. The testing is performed by projecting the query onto the DOVE as well as new representations of semantic concepts from training, with non-negativity and unit summation constraints for probabilistic interpretation. We tested our formulation on TRECVID MED and SIN tasks, and obtained encouraging results.

****

Efficient Feature Extraction, Encoding and Classification for Action Recognition

Vadim Kantorov, Ivan Laptev; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2593-2600

Local video features provide state-of-the-art performance for action recognition. While the accuracy of action recognition has been continuously improved over the recent years, the low speed of feature extraction and subsequent recognition prevents current methods from scaling up to real-size problems. We address this issue and first develop highly efficient video features using motion information in video compression. We next explore feature encoding by Fisher vectors and demonstrate accurate action recognition using fast linear classifiers. Our method improves the speed of video feature extraction, feature encoding and action classification by two orders of magnitude at the cost of minor reduction in recognition accuracy. We validate our approach and compare it to the state of the art on four recent action recognition datasets.

****

3D Pose from Motion for Cross-view Action Recognition via Non-linear Circulant Temporal Encoding

Ankur Gupta, Julieta Martinez, James J. Little, Robert J. Woodham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2601-2608

We describe a new approach to transfer knowledge across views for action recognition by using examples from a large collection of unlabelled mocap data. We achieve this by directly matching purely motion based features from videos to mocap. Our approach recovers 3D pose sequences without performing any body part tracking. We use these matches to generate multiple motion projections and thus add view invariance to our action recognition model. We also introduce a closed form solution for approximate non-linear Circulant Temporal Encoding (nCTE), which allows us to efficiently perform the matches in the frequency domain. We test our approach on the challenging unsupervised modality of the IXMAS dataset, and use publicly available motion capture data for matching. Without any additional annotation effort, we are able to significantly outperform the current state of the art.

****

Human Action Recognition Based on Context-Dependent Graph Kernels

Baoxin Wu, Chunfeng Yuan, Weiming Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2609-2616

Graphs are a powerful tool to model structured objects, but it is nontrivial to measure the similarity between two graphs. In this paper, we construct a two-graph model to represent human actions by recording the spatial and temporal relationships among local features. We also propose a novel family of context-dependent graph kernels (CGKs) to measure similarity between graphs. First, local features are used as the vertices of the two-graph model and the relationships among local features in the intra-frames and inter-frames are characterized by the edges. Then, the proposed CGKs are applied to measure the similarity between actions represented by the two-graph model. Graphs can be decomposed into numbers of primary walk groups with different walk lengths and our CGKs are based on the context-dependent primary walk group matching. Taking advantage of the context information makes the correctly matched primary walk groups dominate in the CGKs and improves the performance of similarity measurement between graphs. Finally, a generalized multiple kernel learning algorithm with a proposed l12-norm regularization is applied to combine these CGKs optimally together and simultaneously train a set of action classifiers. We conduct a series of experiments on several public action datasets. Our approach achieves a comparable performance to the state-of-the-art approaches, which demonstrates the effectiveness of the two-graph model and the CGKs in recognizing human actions.

****

Depth and Skeleton Associated Action Recognition without Online Accessible RGB-D Cameras

Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, Hong-Yuan Mark Liao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2617-2624

The recent advances in RGB-D cameras have allowed us to better solve increasingly complex computer vision tasks. However, modern RGB-D cameras are still restricted by the short effective distances. The limitation may make RGB-D cameras not online accessible in practice, and degrade their applicability. We propose an alternative scenario to address this problem, and illustrate it with the application to action recognition. We use Kinect to offline collect an auxiliary, multi-modal database, in which not only the RGB videos but also the depth maps and skeleton structures of actions of interest are available. Our approach aims to enhance action recognition in RGB videos by leveraging the extra database. Specifically, it optimizes a feature transformation, by which the actions to be recognized can be concisely reconstructed by entries in the auxiliary database. In this way, the inter-database variations are adapted. More importantly, each action can be augmented with additional depth and skeleton images retrieved from the auxiliary database. The proposed approach has been evaluated on three benchmarks of action recognition. The promising results manifest that the augmented depth and skeleton features can lead to remarkable boost in recognition accuracy.

****

DL-SFA: Deeply-Learned Slow Feature Analysis for Action Recognition

Lin Sun, Kui Jia, Tsung-Han Chan, Yuqiang Fang, Gang Wang, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2625-2632

Most of the previous work on video action recognition use complex hand-designed local features, such as SIFT, HOG and SURF, but these approaches are implemented sophisticatedly and difficult to be extended to other sensor modalities. Recent studies discover that there are no universally best hand-engineered features for all datasets, and learning features directly from the data may be more advantageous. One such endeavor is Slow Feature Analysis (SFA) proposed by Wiskott and Sejnowski. SFA can learn the invariant and slowly varying features from input signals and has been proved to be valuable in human action recognition. It is also observed that the multi-layer feature representation has succeeded remarkably in widespread machine learning applications. In this paper, we propose to combine SFA with deep learning techniques to learn hierarchical representations from the video data itself. Specifically, we use a two-layered SFA learning structure with 3D convolution and max pooling operations to scale up the method to large inputs and capture abstract and structural features from the video. Thus, the proposed method is suitable for action recognition. At the same time, sharing the same merits of deep learning, the proposed method is generic and fully automated. Our classification results on Hollywood2, KTH and UCF Sports are competitive with previously published results. To highlight some, on the KTH dataset, our recognition rate shows approximately 1% improvement in comparison to state-of-the-art methods even without supervision or dense sampling.

****

A Cause and Effect Analysis of Motion Trajectories for Modeling Actions

Sanath Narayan, Kalpathi R. Ramakrishnan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2633-2640

An action is typically composed of different parts of the object moving in particular sequences. The presence of different motions (represented as a 1D histogram) has been used in the traditional bag-of-words (BoW) approach for recognizing actions. However the interactions among the motions also form a crucial part of an action. Different object-parts have varying degrees of interactions with the other parts during an action cycle. It is these interactions we want to quantify in order to bring in additional information about the actions. In this paper we propose a causality based approach for quantifying the interactions to aid action classification. Granger causality is used to compute the cause and effect relationships for pairs of motion trajectories of a video. A 2D histogram descriptor for the video is constructed using these pairwise measures. Our proposed method of obtaining pairwise measures for videos is also applicable for large datasets. We have conducted experiments on challenging action recognition databases such as HMDB51 and UCF50 and shown that our causality descriptor helps in encoding additional information regarding the actions and performs on par with the state-of-the art approaches. Due to the complementary nature, a further increase in performance can be observed by combining our approach with state-of-the-art approaches.

****

From Stochastic Grammar to Bayes Network: Probabilistic Parsing of Complex Activity

Nam N. Vo, Aaron F. Bobick; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2641-2648

We propose a probabilistic method for parsing a temporal sequence such as a complex activity defined as composition of sub-activities/actions. The temporal structure of the high-level activity is represented by a string-length limited stochastic context-free grammar. Given the grammar, a Bayes network, which we term Se quential Interval Network (SIN), is generated where the variable nodes correspond to the start and end times of component actions. The network integrates information about the duration of each primitive action, visual detection results for each primitive action, and the activity's temporal structure. At any moment in time during the activity, message passing is used to perform exact inference yielding the posterior probabilities of the start and end times for each different activity/action. We provide demonstrations of this framework being applied to vision tasks such as action prediction, classification of the high-level activities or temporal segmentation of a test sequence; the method is also applicable in Human Robot Interaction domain where continual prediction of human action is needed.

****

Cross-view Action Modeling, Learning and Recognition

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, Song-Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2649-2656

Existing methods on video-based action recognition are generally view-dependent, i.e., performing recognition from the same views seen in the training data. We present a novel multiview spatio-temporal AND-OR graph (MST-AOG) representation for cross-view action recognition, i.e., the recognition is performed on the video from an unknown and unseen view. As a compositional model, MST-AOG compactly represents the hierarchical combinatorial structures of cross-view actions by explicitly modeling the geometry, appearance and motion variations. This paper proposes effective methods to learn the structure and parameters of MST-AOG. The inference based on MST-AOG enables action recognition from novel views. The training of MST-AOG takes advantage of the 3D human skeleton data obtained from Kinect cameras to avoid annotating enormous multi-view video frames, which is error-prone and time-consuming, but the recognition does not need 3D information and is based on 2D video input. A new Multiview Action3D dataset has been created and will be released. Extensive experiments have demonstrated that this new action representation significantly improves the accuracy and robustness for cross-view action recognition on 2D videos.

****

Visual Semantic Search: Retrieving Videos via Complex Textual Queries

Dahua Lin, Sanja Fidler, Chen Kong, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2657-2664

In this paper, we tackle the problem of retrieving videos using complex natural language queries. Towards this goal, we first parse the sentential descriptions into a semantic graph, which is then matched to visual concepts using a generalized bipartite matching algorithm. Our approach exploits object appearance, motion and spatial relations, and learns the importance of each term using structure prediction. We demonstrate the effectiveness of our approach on a new dataset designed for semantic search in the context of autonomous driving, which exhibits complex and highly dynamic scenes with many objects. We show that our approach is able to locate a major portion of the objects described in the query with high accuracy, and improve the relevance in video retrieval.

****

Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, Pradeep Natarajan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2665-2672

Current state-of-the-art systems for visual content analysis require large training sets for each class of interest, and performance degrades rapidly with fewer examples. In this paper, we present a general framework for the zeroshot learning problem of performing high-level event detection with no training exemplars, using only textual descriptions. This task goes beyond the traditional zero-shot framework of adapting a given set of classes with training data to unseen classes. We leverage video and image collections with free-form text descriptions from widely available web sources to learn a large bank of concepts, in addition to using several off-the-shelf concept detectors, speech, and video text for repre senting videos. We utilize natural language processing technologies to generate event description features. The extracted features are then projected to a common high-dimensional space using text expansion, and similarity is computed in this space. We present extensive experimental results on the large TRECVID MED corpus to demonstrate our approach. Our results show that the proposed concept detection methods significantly outperform current attribute classifiers such as Classemes, ObjectBank, and SUN attributes. Further, we find that fusion, both within as well as between modalities, is crucial for optimal performance.

****

Dual Linear Regression Based Classification for Face Cluster Recognition

Liang Chen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2673-2680

We are dealing with the face cluster recognition problem where there are multiple images per subject in both gallery and probe sets. It is never guaranteed to have a clear spatio-temporal relation among the multiple images of each subject. Considering that the image vectors of each subject, either in gallery or in probe, span a subspace; an algorithm, Dual Linear Regression Classification (DLRC), for the face cluster recognition problem is developed where the distance between two subspaces is defined as the similarity value between a gallery subject and a probe subject. DLRC attempts to find a "virtual" face image located in the intersection of the subspaces spanning from both clusters of face images. The "distance" between the "virtual" face images reconstructed from both subspaces is then taken as the distance between these two subspaces. We further prove that such distance can be formulated under a single linear regression model where we indeed can find the "distance" without reconstructing the "virtual" face images. Extensive experimental evaluations demonstrated the effectiveness of DLRC algorithm compared to other algorithms.

****

Bags of Spacetime Energies for Dynamic Scene Recognition

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2681-2688

This paper presents a unified bag of visual word (BoW) framework for dynamic scene recognition. The approach builds on primitive features that uniformly capture spatial and temporal orientation structure of the imagery (e.g., video), as extracted via application of a bank of spatiotemporally oriented filters. Various feature encoding techniques are investigated to abstract the primitives to an intermediate representation that is best suited to dynamic scene representation. Further, a novel approach to adaptive pooling of the encoded features is presented that captures spatial layout of the scene even while being robust to situations where camera motion and scene dynamics are confounded. The resulting overall approach has been evaluated on two standard, publically available dynamic scene datasets. The results show that in comparison to a representative set of alternatives, the proposed approach outperforms the previous state-of-the-art in classification accuracy by 10%.

****

Feature-Independent Action Spotting Without Human Localization, Segmentation or Frame-wise Tracking

Chuan Sun, Marshall Tappen, Hassan Foroosh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2689-2696

In this paper, we propose an unsupervised framework for action spotting in videos that does not depend on any specific feature (e.g. HOG/HOF, STIP, silhouette, bag-of-words, etc.). Furthermore, our solution requires no human localization, segmentation, or framewise tracking. This is achieved by treating the problem holistically as that of extracting the internal dynamics of video cuboids by modeling them in their natural form as multilinear tensors. To extract their internal dynamics, we devised a novel Two-Phase Decomposition (TP-Decomp) of a tensor that generates very compact and discriminative representations that are robust to even heavily perturbed data. Technically, a Rank-based Tensor Core Pyramid (Rank-TCP) descriptor is generated by combining multiple tensor cores under multiple ranks, allowing to represent video cuboids in a hierarchical tensor pyramid. The problem then reduces to a template matching problem, which is solved efficiently by using two boosting strategies: (1) to reduce search space, we filter the dense trajectory cloud extracted from the target video; (2) to boost the matching speed, we perform matching in an iterative coarse-to-fine manner. Experiments on 5 benchmarks show that our method outperforms current state-of-the-art under various challenging conditions. We also created a challenging dataset called Heavily Perturbed Video Array (HPVA) to validate the robustness of our framework under heavily perturbed situations.

****

Multiscale Centerline Detection by Learning a Scale-Space Distance Transform

Amos Sironi, Vincent Lepetit, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2697-2704

We propose a robust and accurate method to extract the centerlines and scale of tubular structures in 2D images and 3D volumes. Existing techniques rely either on filters designed to respond to ideal cylindrical structures, which lose accuracy when the linear structures become very irregular, or on classification, which is inaccurate because locations on centerlines and locations immediately next to them are extremely difficult to distinguish. We solve this problem by reformulating centerline detection in terms of a regression problem. We first train regressors to return the distances to the closest centerline in scale-space, and we apply them to the input images or volumes. The centerlines and the corresponding scale then correspond to the regressors local maxima, which can be easily identified. We show that our method outperforms state-of-the-art techniques for various 2D and 3D datasets.

****

Multivariate General Linear Models (MGLM) on Riemannian Manifolds with Applications to Statistical Analysis of Diffusion Weighted Images

Hyunwoo J. Kim, Nagesh Adluru, Maxwell D. Collins, Moo K. Chung, Barbara B. Bendlin, Sterling C. Johnson, Richard J. Davidson, Vikas Singh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2705-2712

Linear regression is a parametric model which is ubiquitous in scientific analysis. The classical setup where the observations and responses, i.e., (xi,yi) pairs, are Euclidean is well studied. The setting where yi is manifold valued is a topic of much interest, motivated by applications in shape analysis, topic modeling, and medical imaging. Recent work gives strategies for max-margin classifiers, principal components analysis, and dictionary learning on certain types of manifolds. For parametric regression specifically, results within the last year provide mechanisms to regress one real-valued parameter, xi in R, against a manifold-valued variable, yi in M. We seek to substantially extend the operating range of such methods by deriving schemes for multivariate multiple linear regression â■■ a manifold-valued dependent variable against multiple independent variables, i.e., f : Rn -> M. Our variational algorithm efficiently solves for multiple geodesic bases on the manifold concurrently via gradient updates. This allows us to answer questions such as: what is the relationship of the measurement at voxel y to disease when conditioned on age and gender. We show applications to statistical analysis of diffusion weighted images, which give rise to regression tasks on the manifold GL(n)/O(n) for diffusion tensor images (DTI) and the Hilbert unit sphere for orientation distribution functions (ODF) from high angular resolution acquisition. The companion open-source code is available on nitrc.org/projects/riem_mglm.

****

Preconditioning for Accelerated Iteratively Reweighted Least Squares in Structured Sparsity Reconstruction

Chen Chen, Junzhou Huang, Lei He, Hongsheng Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2713-2720

In this paper, we propose a novel algorithm for structured sparsity reconstruction. This algorithm is based on the iterative reweighted least squares (IRLS) framework, and accelerated by the preconditioned conjugate gradient method. The convergence rate of the proposed algorithm is almost the same as that of the tradi tional IRLS algorithms, that is, exponentially fast. Moreover, with the devised preconditioner, the computational cost for each iteration is significantly less than that of traditional IRLS algorithms, which makes it feasible for large scale problems. Besides the fast convergence, this algorithm can be flexibly applied to standard sparsity, group sparsity, and overlapping group sparsity problems. Experiments are conducted on a practical application compressive sensing magnetic resonance imaging. Results demonstrate that the proposed algorithm achieves superior performance over 9 state-of-the-art algorithms in terms of both accuracy and computational cost.

****

Joint Coupled-Feature Representation and Coupled Boosting for AD Diagnosis

Yinghuan Shi, Heung-Il Suk, Yang Gao, Dinggang Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2721-2728

Recently, there has been a great interest in computer-aided Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) diagnosis. Previous learning based methods defined the diagnosis process as a classification task and directly used the low-level features extracted from neuroimaging data without considering relations among them. However, from a neuroscience point of view, it's well known that a human brain is a complex system that multiple brain regions are anatomically connected and functionally inter- act with each other. Therefore, it is natural to hypothesize that the low-level features extracted from neuroimaging data are related to each other in some ways. To this end, in this paper, we first devise a coupled feature representation by utilizing intra-coupled and inter-coupled interaction relationship. Regarding multi-modal data fusion, we propose a novel coupled boosting algorithm that analyzes the pairwise coupled-diversity correlation between modalities. Specifically, we formulate a new weight updating function, which considers both incorrectly and inconsistently classified samples. In our experiments on the ADNI dataset, the proposed method presented the best performance with accuracies of 94.7% and 80.1% for AD vs. Normal Control (NC) and MCI vs. NC classifications, respectively, outperforming the competing methods and the state-of-the-art methods.

****

Deformable Registration of Feature-Endowed Point Sets Based on Tensor Fields

Demian Wassermann, James Ross, George Washko, William M. Wells III, Raul San Jose-Estepar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2729-2735

The main contribution of this work is a framework to register anatomical structures characterized as a point set where each point has an associated symmetric matrix. These matrices can represent problem-dependent characteristics of the registered structure. For example, in airways, matrices can represent the orientation and thickness of the structure. Our framework relies on a dense tensor field representation which we implement sparsely as a kernel mixture of tensor fields. We equip the space of tensor fields with a norm that serves as a similarity measure. To calculate the optimal transformation between two structures we minimize this measure using an analytical gradient for the similarity measure and the deformation field, which we restrict to be a diffeomorphism. We illustrate the value of our tensor field model by comparing our results with scalar and vector field based models. Finally, we evaluate our registration algorithm on synthetic data sets and validate our approach on manually annotated airway trees.

****

Tracking Indistinguishable Translucent Objects over Time using Weakly Supervised Structured Learning

Luca Fiaschi, Ferran Diego, Konstantin Gregor, Martin Schiegg, Ullrich Koethe, Marta Zlatic, Fred A. Hamprecht; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2736-2743

We use weakly supervised structured learning to track and disambiguate the identity of multiple indistinguishable, translucent and deformable objects that can overlap for many frames. For this challenging problem, we propose a novel model which handles occlusions, complex motions and non-rigid deformations by jointly optimizing the flows of multiple latent intensities across frames. These flows ar e latent variables for which the user cannot directly provide labels. Instead, we leverage a structured learning formulation that uses weak user annotations to find the best hyperparameters of this model. The approach is evaluated on a challenging dataset for the tracking of multiple Drosophila larvae which we make publicly available. Our method tracks multiple larvae in spite of their poor distinguishability and minimizes the number of identity switches during prolonged mutual occlusion.

****

Scale-space Processing Using Polynomial Representations

Gou Koutaki, Keiichi Uchimura; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2744-2751

In this study, we propose the application of principal components analysis (PCA) to scale-spaces. PCA is a standard method used in computer vision. The translation of an input image into scale-space is a continuous operation, which requires the extension of conventional finite matrix-based PCA to an infinite number of dimensions. In this study, we use spectral decomposition to resolve this infinite eigenproblem by integration and we propose an approximate solution based on polynomial equations. To clarify its eigensolutions, we apply spectral decomposition to the Gaussian scale-space and scale-normalized Laplacian of Gaussian (LoG) space. As an application of this proposed method, we introduce a method for generating Gaussian blur images and scale-normalized LoG images, where we demonstrate that the accuracy of these images can be very high when calculating an arbitrary scale using a simple linear combination. We also propose a new Scale Invariant Feature Transform (SIFT) detector as a more practical example.

****

Single Image Layer Separation using Relative Smoothness

Yu Li, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2752-2759

This paper addresses extracting two layers from an image where one layer is smoother than the other. This problem arises most notably in intrinsic image decomposition and reflection interference removal. Layer decomposition from a single-image is inherently ill-posed and solutions require additional constraints to be enforced. We introduce a novel strategy that regularizes the gradients of the two layers such that one has a long tail distribution and the other a short tail distribution. While imposing the long tail distribution is a common practice, our introduction of the short tail distribution on the second layer is unique. We formulate our problem in a probabilistic framework and describe an optimization scheme to solve this regularization with only a few iterations. We apply our approach to the intrinsic image and reflection removal problems and demonstrate high quality layer separation on par with other techniques but being significantly faster than prevailing methods.

****

Image Fusion with Local Spectral Consistency and Dynamic Gradient Sparsity

Chen Chen, Yeqing Li, Wei Liu, Junzhou Huang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2760-2765

In this paper, we propose a novel method for image fusion from a high resolution panchromatic image and a low resolution multispectral image at the same geographical location. Different from previous methods, we do not make any assumption about the upsampled multispectral image, but only assume that the fused image after downsampling should be close to the original multispectral image. This is a severely ill-posed problem and a dynamic gradient sparsity penalty is thus proposed for regularization. Incorporating the intra- correlations of different bands, this penalty can effectively exploit the prior information (e.g. sharp boundaries) from the panchromatic image. A new convex optimization algorithm is proposed to efficiently solve this problem. Extensive experiments on four multispectral datasets demonstrate that the proposed method significantly outperforms the state-of-the-arts in terms of both spatial and spectral qualities.

****

Segmentation-Free Dynamic Scene Deblurring

Tae Hyun Kim, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Visi on and Pattern Recognition (CVPR), 2014, pp. 2766-2773

Most state-of-the-art dynamic scene deblurring methods based on accurate motion segmentation assume that motion blur is small or that the specific type of motion causing the blur is known. In this paper, we study a motion segmentation-free dynamic scene deblurring method, which is unlike other conventional methods. When the motion can be approximated to linear motion that is locally (pixel-wise) varying, we can handle various types of blur caused by camera shake, including out-of-plane motion, depth variation, radial distortion, and so on. Thus, we propose a new energy model simultaneously estimating motion flow and the latent image based on robust total variation (TV)-L1 model. This approach is necessary to handle abrupt changes in motion without segmentation. Furthermore, we address the problem of the traditional coarse-to-fine deblurring framework, which gives rise to artifacts when restoring small structures with distinct motion. We thus propose a novel kernel re-initialization method which reduces the error of motion flow propagated from a coarser level. Moreover, a highly effective convex optimization-based solution mitigating the computational difficulties of the TV-L1 model is established. Comparative experimental results on challenging real blurry images demonstrate the efficiency of the proposed method.

****

Shrinkage Fields for Effective Image Restoration

Uwe Schmidt, Stefan Roth; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2774-2781

Many state-of-the-art image restoration approaches do not scale well to larger images, such as megapixel images common in the consumer segment. Computationally expensive optimization is often the culprit. While efficient alternatives exist, they have not reached the same level of image quality. The goal of this paper is to develop an effective approach to image restoration that offers both computational efficiency and high restoration quality. To that end we propose shrinkage fields, a random field-based architecture that combines the image model and the optimization algorithm in a single unit. The underlying shrinkage operation bears connections to wavelet approaches, but is used here in a random field context. Computational efficiency is achieved by construction through the use of convolution and DFT as the core components; high restoration quality is attained through loss-based training of all model parameters and the use of a cascade architecture. Unlike heavily engineered solutions, our learning approach can be adapted easily to different trade-offs between efficiency and image quality. We demonstrate state-of-the-art restoration results with high levels of computational efficiency, and significant speedup potential through inherent parallelism.

****

Camouflaging an Object from Many Viewpoints

Andrew Owens, Connelly Barnes, Alex Flint, Hanumant Singh, William Freeman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2782-2789

We address the problem of camouflaging a 3D object from the many viewpoints that one might see it from. Given photographs of an object's surroundings, we produce a surface texture that will make the object difficult for a human to detect. To do this, we introduce several background matching algorithms that attempt to make the object look like whatever is behind it. Of course, it is impossible to exactly match the background from every possible viewpoint. Thus our models are forced to make trade-offs between different perceptual factors, such as the conspicuousness of the occlusion boundaries and the amount of texture distortion. We use experiments with human subjects to evaluate the effectiveness of these models for the task of camouflaging a cube, finding that they significantly outperform naÃ¯ve strategies.

****

Learning Optimal Seeds for Diffusion-based Salient Object Detection

Song Lu, Vijay Mahadevan, Nuno Vasconcelos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2790-2797

In diffusion-based saliency detection, an image is partitioned into superpixels and mapped to a graph, with superpixels as nodes and edge strengths proportional to superpixel similarity. Saliency information is then propagated over the graph using a diffusion process, whose equilibrium state yields the object saliency map. The optimal solution is the product of a propagation matrix and a saliency seed vector that contains a prior saliency assessment. This is obtained from either a bottom-up saliency detector or some heuristics. In this work, we propose a method to learn optimal seeds for object saliency. Two types of features are computed per superpixel: the bottom-up saliency of the superpixel region and a set of mid-level vision features informative of how likely the superpixel is to belong to an object. The combination of features that best discriminates between object and background saliency is then learned, using a large-margin formulation of the discriminant saliency principle. The propagation of the resulting saliency seeds, using a diffusion process, is finally shown to outperform the state of the art on a number of salient object detection datasets.

****

Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images

Eleonora Vig, Michael Dorr, David Cox; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2798-2805

Saliency prediction typically relies on hand-crafted (multiscale) features that are combined in different ways to form a "master" saliency map, which encodes local image conspicuity. Recent improvements to the state of the art on standard benchmarks such as MIT1003 have been achieved mostly by incrementally adding more and more hand-tuned features (such as car or face detectors) to existing models. In contrast, we here follow an entirely automatic data-driven approach that performs a large-scale search for optimal features. We identify those instances of a richly-parameterized bio-inspired model family (hierarchical neuromorphic networks) that successfully predict image saliency. Because of the high dimensionality of this parameter space, we use automated hyperparameter optimization to efficiently guide the search. The optimal blend of such multilayer features combined with a simple linear classifier achieves excellent performance on several image saliency benchmarks. Our models outperform the state of the art on MIT1003, on which features and classifiers are learned. Without additional training, these models generalize well to two other image saliency data sets, Toronto and NUSEF, despite their different image content. Finally, our algorithm scores best of all the 23 models evaluated to date on the MIT300 saliency challenge, which uses a hidden test set to facilitate an unbiased comparison.

****

Saliency Detection on Light Field

Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2806-2813

Existing saliency detection approaches use images as inputs and are sensitive to foreground/background similarities, complex background textures, and occlusions. We explore the problem of using light fields as input for saliency detection. Our technique is enabled by the availability of commercial plenoptic cameras that capture the light field of a scene in a single shot. We show that the unique refocusing capability of light fields provides useful focusness, depths, and objectness cues. We further develop a new saliency detection algorithm tailored for light fields. To validate our approach, we acquire a light field database of a range of indoor and outdoor scenes and generate the ground truth saliency map. Experiments show that our saliency detection scheme can robustly handle challenging scenarios such as similar foreground and background, cluttered background, complex occlusions, etc., and achieve high accuracy and robustness.

****

Saliency Optimization from Robust Background Detection

Wangjiang Zhu, Shuang Liang, Yichen Wei, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2814-2821

Recent progresses in salient object detection have exploited the boundary prior, or background information, to assist other saliency cues such as contrast, achieving state-of-the-art results. However, their usage of boundary prior is very simple, fragile, and the integration with other cues is mostly heuristic. In this work, we present new methods to address these issues. First, we propose a robust background measure, called boundary connectivity. It characterizes the spatial layout of image regions with respect to image boundaries and is much more robust. It has an intuitive geometrical interpretation and presents unique benefits that are absent in previous saliency measures. Second, we propose a principled optimization framework to integrate multiple low level cues, including our background measure, to obtain clean and uniform saliency maps. Our formulation is intuitive, efficient and achieves state-of-the-art results on several benchmark datasets.

****

A Reverse Hierarchy Model for Predicting Eye Fixations

Tianlin Shi, Ming Liang, Xiaolin Hu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2822-2829

A number of psychological and physiological evidences suggest that early visual attention works in a coarse-to-fine way, which lays a basis for the reverse hierarchy theory (RHT). This theory states that attention propagates from the top level of the visual hierarchy that processes gist and abstract information of input, to the bottom level that processes local details. Inspired by the theory, we develop a computational model for saliency detection in images. First, the original image is downsampled to different scales to constitute a pyramid. Then, saliency on each layer is obtained by image super-resolution reconstruction from the layer above, which is defined as unpredictability from this coarse-to-fine reconstruction. Finally, saliency on each layer of the pyramid is fused into stochastic fixations through a probabilistic model, where attention initiates from the top layer and propagates downward through the pyramid. Extensive experiments on two standard eye-tracking datasets show that the proposed method can achieve competitive results with state-of-the-art models.

****

100+ Times Faster Weighted Median Filter (WMF)

Qi Zhang, Li Xu, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2830-2837

Weighted median, in the form of either solver or filter, has been employed in a wide range of computer vision solutions for its beneficial properties in sparsity representation. But it is hard to be accelerated due to the spatially varying weight and the median property. We propose a few efficient schemes to reduce computation complexity from O(r^2) to O(r) where r is the kernel size. Our contribution is on a new joint-histogram representation, median tracking, and a new data structure that enables fast data access. The effectiveness of these schemes is demonstrated on optical flow estimation, stereo matching, structure-texture separation, image filtering, to name a few. The running time is largely shortened from several minutes to less than 1 second. The source code is provided in the project website.

****

Edge-aware Gradient Domain Optimization Framework for Image Filtering by Local Propagation

Miao Hua, Xiaohui Bie, Minying Zhang, Wencheng Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2838-2845

Gradient domain methods are popular for image processing. However, these methods even the edge-preserving ones cannot preserve edges well in some cases. In this paper, we present new constraints explicitly to better preserve edges for general gradient domain image filtering and theoretically analyse why these constraints are edge-aware. Our edge-aware constraints are easy to implement, fast to compute and can be seamlessly integrated into the general gradient domain optimization framework. The improved framework can better preserve edges while maintaining similar image filtering effects as the original image filters. We also demonstrate the strength of our edge-aware constraints on various applications such as image smoothing, image colorization and Poisson image cloning.

****

Super-Resolving Noisy Images

Abhishek Singh, Fatih Porikli, Narendra Ahuja; Proceedings of the IEEE Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2846-2853

Our goal is to obtain a noise-free, high resolution (HR) image, from an observed, noisy, low resolution (LR) image. The conventional approach of preprocessing the image with a denoising algorithm, followed by applying a super-resolution (SR) algorithm, has an important limitation: Along with noise, some high frequency content of the image (particularly textural detail) is invariably lost during the denoising step. This 'denoising loss' restricts the performance of the subsequent SR step, wherein the challenge is to synthesize such textural details. In this paper, we show that high frequency content in the noisy image (which is ordinarily removed by denoising algorithms) can be effectively used to obtain the missing textural details in the HR domain. To do so, we first obtain HR versions of both the noisy and the denoised images, using a patch-similarity based SR algorithm. We then show that by taking a convex combination of orientation and frequency selective bands of the noisy and the denoised HR images, we can obtain a desired HR image where (i) some of the textural signal lost in the denoising step is effectively recovered in the HR domain, and (ii) additional textures can be easily synthesized by appropriately constraining the parameters of the convex combination. We show that this part-recovery and part-synthesis of textures through our algorithm yields HR images that are visually more pleasing than those obtained using the conventional processing pipeline. Furthermore, our results show a consistent improvement in numerical metrics, further corroborating the ability of our algorithm to recover lost signal.

****

Sparse Dictionary Learning for Edit Propagation of High-Resolution Images

Xiaowu Chen, Dongqing Zou, Jianwei Li, Xiaochun Cao, Qinping Zhao, Hao Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2854-2861

We introduce a method of sparse dictionary learning for edit propagation of high-resolution images or video. Previous approaches for edit propagation typically employ a global optimization over the whole set of image pixels, incurring a prohibitively high memory and time consumption for high-resolution images. Rather than propagating an edit pixel by pixel, we follow the principle of sparse representation to obtain a compact set of representative samples (or features) and perform edit propagation on the samples instead. The sparse set of samples provides an intrinsic basis for an input image, and the coding coefficients capture the linear relationship between all pixels and the samples. The representative set of samples is then optimized by a novel scheme which maximizes the KL-divergence between each sample pair to remove redundant samples. We show several applications of sparsity-based edit propagation including video recoloring, theme editing, and seamless cloning, operating on both color and texture features. We demonstrate that with a sample-to-pixel ratio in the order of 0.01%, signifying a significant reduction on memory consumption, our method still maintains a high-degree of visual fidelity.

****

Weighted Nuclear Norm Minimization with Application to Image Denoising

Shuhang Gu, Lei Zhang, Wangmeng Zuo, Xiangchu Feng; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2862-2869

As a convex relaxation of the low rank matrix factorization problem, the nuclear norm minimization has been attracting significant research interest in recent years. The standard nuclear norm minimization regularizes each singular value equally to pursue the convexity of the objective function. However, this greatly restricts its capability and flexibility in dealing with many practical problems (e.g., denoising), where the singular values have clear physical meanings and should be treated differently. In this paper we study the weighted nuclear norm minimization (WNNM) problem, where the singular values are assigned different weights. The solutions of the WNNM problem are analyzed under different weighting conditions. We then apply the proposed WNNM algorithm to image denoising by exploiting the image nonlocal self-similarity. Experimental results clearly show that the proposed WNNM algorithm outperforms many state-of-the-art denoising algorithms such as BM3D in terms of both quantitative measure and visual perception quali ty.

****

Using Projection Kurtosis Concentration Of Natural Images For Blind Noise Covariance Matrix Estimation

Xing Zhang, Siwei Lyu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2870-2876

Kurtosis of 1D projections provides important statistical characteristics of natural images. In this work, we first provide a theoretical underpinning to a recently observed phenomenon known as projection kurtosis concentration that the kurtosis of natural images over different band-pass channels tend to concentrate around a typical value. Based on this analysis, we further describe a new method to estimate the covariance matrix of correlated Gaussian noise from a noise corrupted image using random band-pass filters. We demonstrate the effectiveness of our blind noise covariance matrix estimation method on natural images.

****

Blind Image Quality Assessment using Semi-supervised Rectifier Networks

Huixuan Tang, Neel Joshi, Ashish Kapoor; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2877-2884

It is often desirable to evaluate images quality with a perceptually relevant measure that does not require a reference image. Recent approaches to this problem use human provided quality scores with machine learning to learn a measure. The biggest hurdles to these efforts are: 1) the difficulty of generalizing across diverse types of distortions and 2) collecting the enormity of human scored training data that is needed to learn the measure. We present a new blind image quality measure that addresses these difficulties by learning a robust, nonlinear kernel regression function using a rectifier neural network. The method is pre-trained with unlabeled data and fine-tuned with labeled data. It generalizes across a large set of images and distortion types without the need for a large amount of labeled data. We evaluate our approach on two benchmark datasets and show that it not only outperforms the current state of the art in blind image quality estimation, but also outperforms the state of the art in non-blind measures. Furthermore, we show that our semi-supervised approach is robust to using varying amounts of labeled data.

****

Separable Kernel for Image Deblurring

Lu Fang, Haifeng Liu, Feng Wu, Xiaoyan Sun, Houqiang Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2885-2892

In this paper, we deal with the image deblurring problem in a completely new perspective by proposing separable kernel to represent the inherent properties of the camera and scene system. Specifically, we decompose a blur kernel into three individual descriptors (trajectory, intensity and point spread function) so that they can be optimized separately. To demonstrate the advantages, we extract one-pixel-width trajectories of blur kernels and propose a random perturbation algorithm to optimize them but still keeping their continuity. For many cases, where current deblurring approaches fall into local minimum, excellent deblurred results and correct blur kernels can be obtained by individually optimizing the kernel trajectories. Our work strongly suggests that more constraints and priors should be introduced to blur kernels in solving the deblurring problem because blur kernels have lower dimensions than images.

****

Joint Depth Estimation and Camera Shake Removal from Single Blurry Image

Zhe Hu, Li Xu, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2893-2900

Camera shake during exposure time often results in spatially variant blur effect of the image. The non-uniform blur effect is not only caused by the camera motion, but also the depth variation of the scene. The objects close to the camera sensors are likely to appear more blurry than those at a distance in such cases. However, recent non-uniform deblurring methods do not explicitly consider the depth factor or assume fronto-parallel scenes with constant depth for simplicity.

While single image non-uniform deblurring is a challenging problem, the blurry results in fact contain depth information which can be exploited. We propose to jointly estimate scene depth and remove non-uniform blur caused by camera motion by exploiting their underlying geometric relationships, with only single blurry image as input. To this end, we present a unified layer-based model for depth-involved deblurring. We provide a novel layer-based solution using matting to partition the layers and an expectation-maximization scheme to solve this problem. This approach largely reduces the number of unknowns and makes the problem tractable. Experiments on challenging examples demonstrate that both depth and camera shake removal can be well addressed within the unified framework.

****

Deblurring Text Images via L0-Regularized Intensity and Gradient Prior

Jinshan Pan, Zhe Hu, Zhixun Su, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2901-2908

We propose a simple yet effective L_0-regularized prior based on intensity and gradient for text image deblurring. The proposed image prior is motivated by observing distinct properties of text images. Based on this prior, we develop an efficient optimization method to generate reliable intermediate results for kernel estimation. The proposed method does not require any complex filtering strategies to select salient edges which are critical to the state-of-the-art deblurring algorithms. We discuss the relationship with other deblurring algorithms based on edge selection and provide insight on how to select salient edges in a more principled way. In the final latent image restoration step, we develop a simple method to remove artifacts and render better deblurred images. Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art text image deblurring methods. In addition, we show that the proposed method can be effectively applied to deblur low-illumination images.

****

Total Variation Blind Deconvolution: The Devil is in the Details

Daniele Perrone, Paolo Favaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2909-2916

In this paper we study the problem of blind deconvolution. Our analysis is based on the algorithm of Chan and Wong [2] which popularized the use of sparse gradient priors via total variation. We use this algorithm because many methods in the literature are essentially adaptations of this framework. Such algorithm is an iterative alternating energy minimization where at each step either the sharp image or the blur function are reconstructed. Recent work of Levin et al. [14] showed that any algorithm that tries to minimize that same energy would fail, as the desired solution has a higher energy than the no-blur solution, where the sharp image is the blurry input and the blur is a Dirac delta. However, experimentally one can observe that Chan and Wong's algorithm converges to the desired solution even when initialized with the no-blur one. We provide both analysis and experiments to resolve this paradoxical conundrum. We find that both claims are right. The key to understanding how this is possible lies in the details of Chan and Wong's implementation and in how seemingly harmless choices result in dramatic effects. Our analysis reveals that the delayed scaling (normalization) in the iterative step of the blur kernel is fundamental to the convergence of the algorithm. This then results in a procedure that eludes the no-blur solution, despite it being a global minimum of the original energy. We introduce an adaptation of this algorithm and show that, in spite of its extreme simplicity, it is very robust and achieves a performance comparable to the state of the art.

****

Single Image Super-resolution using Deformable Patches

Yu Zhu, Yanning Zhang, Alan L. Yuille; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2917-2924

We proposed a deformable patches based method for single image super-resolution. By the concept of deformation, a patch is not regarded as a fixed vector but a flexible deformation flow. Via deformable patches, the dictionary can cover more patterns that do not appear, thus becoming more expressive. We present the energy function with slow, smooth and flexible prior for deformation model. During e

xample-based super-resolution, we develop the deformation similarity based on the minimized energy function for basic patch matching. For robustness, we utilize multiple deformed patches combination for the final reconstruction. Experiments evaluate the deformation effectiveness and super-resolution performance, showing that the deformable patches help improve the representation accuracy and perform better than the state-of-art methods.
****
Multi-Shot Imaging: Joint Alignment, Deblurring and Resolution-Enhancement
Haichao Zhang, Lawrence Carin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2925-2932
The capture of multiple images is a simple way to increase the chance of capturing a good photo with a light-weight hand-held camera, for which the camera-shake blur is typically a nuisance problem. The naive approach of selecting the single best captured photo as output does not take full advantage of all the observations. Conventional multi-image blind deblurring methods can take all observations as input but usually require the multiple images are well aligned. However, the multiple blurry images captured in the presence of camera shake are rarely free from mis-alignment. Registering multiple blurry images is a challenging task due to the presence of blur while deblurring of multiple blurry images requires accurate alignment, leading to an intrinsically coupled problem. In this paper, we propose a blind multi-image restoration method which can achieve joint alignment, non-uniform deblurring, together with resolution enhancement from multiple low-quality images. Experiments on several real-world images with comparison to some previous methods validate the effectiveness of the proposed method.
****
CID: Combined Image Denoising in Spatial and Frequency Domains Using Web Images
Huanjing Yue, Xiaoyan Sun, Jingyu Yang, Feng Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2933-2940
In this paper, we propose a novel two-step scheme to filter heavy noise from images with the assistance of retrieved Web images. There are two key technical contributions in our scheme. First, for every noisy image block, we build two three dimensional (3D) data cubes by using similar blocks in retrieved Web images and similar nonlocal blocks within the noisy image, respectively. To better use their correlations, we propose different denoising strategies. The denoising in the 3D cube built upon the retrieved images is performed as median filtering in the spatial domain, whereas the denoising in the other 3D cube is performed in the frequency domain. These two denoising results are then combined in the frequency domain to produce a denoising image. Second, to handle heavy noise, we further propose using the denoising image to improve image registration of the retrieved Web images, 3D cube building, and the estimation of filtering parameters in the frequency domain. Afterwards, the proposed denoising is performed on the noisy image again to generate the final denoising result. Our experimental results show that when the noise is high, the proposed scheme is better than BM3D by more than 2 dB in PSNR and the visual quality improvement is clear to see.
****
Multipoint Filtering with Local Polynomial Approximation and Range Guidance
Xiao Tan, Changming Sun, Tuan D. Pham; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2941-2948
This paper presents a novel guided image filtering method using multipoint local polynomial approximation (LPA) with range guidance. In our method, the LPA is extended from a pointwise model into a multipoint model for reliable filtering and better preserving image spatial variation which usually contains the essential information in the input image. In addition, we develop a scheme with constant computational complexity (invariant to the size of filtering kernel) for generating a spatial adaptive support region around a point. By using the hybrid of the local polynomial model and color/intensity based range guidance, the proposed method not only preserves edges but also does a much better job in preserving spatial variation than existing popular filtering methods. Our method proves to be effective in a number of applications: depth image upsampling, joint image denoising, details enhancement, and image abstraction. Experimental results show tha

t our method produces better results than state-of-the-art methods and it is also computationally efficient.

****

Decomposable Nonlocal Tensor Dictionary Learning for Multispectral Image Denoising

Yi Peng, Deyu Meng, Zongben Xu, Chenqiang Gao, Yi Yang, Biao Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2949-2956

As compared to the conventional RGB or gray-scale images, multispectral images (MSI) can deliver more faithful representation for real scenes, and enhance the performance of many computer vision tasks. In practice, however, an MSI is always corrupted by various noises. In this paper we propose an effective MSI denoising approach by combinatorially considering two intrinsic characteristics underlying an MSI: the nonlocal similarity over space and the global correlation across spectrum. In specific, by explicitly considering spatial self-similarity of an MSI we construct a nonlocal tensor dictionary learning model with a group-block-sparsity constraint, which makes similar full-band patches (FBP) share the same atoms from the spatial and spectral dictionaries. Furthermore, through exploiting spectral correlation of an MSI and assuming over-redundancy of dictionaries, the constrained nonlocal MSI dictionary learning model can be decomposed into a series of unconstrained low-rank tensor approximation problems, which can be readily solved by off-the-shelf higher order statistics. Experimental results show that our method outperforms all state-of-the-art MSI denoising methods under comprehensive quantitative performance measures.

****

Robust 3D Features for Matching between Distorted Range Scans Captured by Moving Systems

Xiangqi Huang, Bo Zheng, Takeshi Masuda, Katsushi Ikeuchi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2957-2964

Laser range sensors are often demanded to mount on a moving platform for achieving the good efficiency of 3D reconstruction. However, such moving systems often suffer from the difficulty of matching the distorted range scans. In this paper, we propose novel 3D features which can be robustly extracted and matched even for the distorted 3D surface captured by a moving system. Our feature extraction employs Morse theory to construct Morse functions which capture the critical points approximately invariant to the 3D surface distortion. Then for each critical point, we extract support regions with the maximally stable region defined by extremal region or disconnectivity. Our feature description is designed as two steps: 1) we normalize the detected local regions to canonical shapes for robust matching; 2) we encode each key point with multiple vectors at different Morse function values. In experiments, we demonstrate that the proposed 3D features achieve substantially better performance for distorted surface matching than the state-of-the-art methods.

****

Discriminative Blur Detection Features

Jianping Shi, Li Xu, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2965-2972

Ubiquitous image blur brings out a practically important question â■■ what are effective features to differentiate between blurred and unblurred image regions. We address it by studying a few blur feature representations in image gradient, Fourier domain, and data-driven local filters. Unlike previous methods, which are often based on restoration mechanisms, our features are constructed to enhance discriminative power and are adaptive to various blur scales in images. To avail evaluation, we build a new blur perception dataset containing thousands of images with labeled ground-truth. Our results are applied to several applications, including blur region segmentation, deblurring, and blur magnification.

****

Detection, Rectification and Segmentation of Coplanar Repeated Patterns

James Pritts, Ondrej Chum, Jiri Matas; Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition (CVPR), 2014, pp. 2973-2980

This paper presents a novel and general method for the detection, rectification and segmentation of imaged coplanar repeated patterns. The only assumption made of the scene geometry is that repeated scene elements are mapped to each other by planar Euclidean transformations. The class of patterns covered is broad and includes nearly all commonly seen, planar, man-made repeated patterns. In addition, novel linear constraints are used to reduce geometric ambiguity between the rectified imaged pattern and the scene pattern. Rectification to within a similarity of the scene plane is achieved from one rotated repeat, or to within a similarity with a scale ambiguity along the axis of symmetry from one reflected repeat. A stratum of constraints is derived that gives the necessary configuration of repeats for each successive level of rectification. A generative model for the imaged pattern is inferred and used to segment the pattern with pixel accuracy. Qualitative results are shown on a broad range of image types on which state-of-the-art methods fail.

****

Mirror Symmetry Histograms for Capturing Geometric Properties in Images

Marcelo Cicconet, Davi Geiger, Kristin C. Gunsalus, Michael Werman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2981-2986

We propose a data structure that captures global geometric properties in images: Histogram of Mirror Symmetry Coefficients. We compute such a coefficient for every pair of pixels, and group them in a 6-dimensional histogram. By marginalizing the HMSC in various ways, we develop algorithms for a range of applications: detection of nearly-circular cells; location of the main axis of reflection symmetry; detection of cell-division in movies of developing embryos; detection of worm-tips and indirect cell-counting via supervised classification. Our approach generalizes a series of histogram-related methods, and the proposed algorithms perform with state-of-the-art accuracy.

****

A Learning-to-Rank Approach for Image Color Enhancement

Jianzhou Yan, Stephen Lin, Sing Bing Kang, Xiaoou Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2987-2994

We present a machine-learned ranking approach for automatically enhancing the color of a photograph. Unlike previous techniques that train on pairs of images before and after adjustment by a human user, our method takes into account the intermediate steps taken in the enhancement process, which provide detailed information on the person's color preferences. To make use of this data, we formulate the color enhancement task as a learning-to-rank problem in which ordered pairs of images are used for training, and then various color enhancements of a novel input image can be evaluated from their corresponding rank values. From the parallels between the decision tree structures we use for ranking and the decisions made by a human during the editing process, we posit that breaking a full enhancement sequence into individual steps can facilitate training. Our experiments show that this approach compares well to existing methods for automatic color enhancement.

****

Investigating Haze-relevant Features in A Learning Framework for Image Dehazing

Ketan Tang, Jianchao Yang, Jue Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2995-3000

Haze is one of the major factors that degrade outdoor images. Removing haze from a single image is known to be severely ill-posed, and assumptions made in previous methods do not hold in many situations. In this paper, we systematically investigate different haze-relevant features in a learning framework to identify the best feature combination for image dehazing. We show that the dark-channel feature is the most informative one for this task, which confirms the observation of He et al. [8] from a learning perspective, while other haze-relevant features also contribute significantly in a complementary way. We also find that surprisingly, the synthetic hazy image patches we use for feature investigation serve we

ll as training data for realworld images, which allows us to train specific models for specific applications. Experiment results demonstrate that the proposed algorithm outperforms state-of-the-art methods on both synthetic and real-world datasets.
****

Quality Assessment for Comparing Image Enhancement Algorithms
Zhengying Chen, Tingting Jiang, Yonghong Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3003-3010
As the image enhancement algorithms developed in recent years, how to compare the performances of different image enhancement algorithms becomes a novel task. In this paper, we propose a framework to do quality assessment for comparing image enhancement algorithms. Not like traditional image quality assessment approaches, we focus on the relative quality ranking between enhanced images rather than giving an absolute quality score for a single enhanced image. We construct a dataset which contains source images in bad visibility and their enhanced images processed by different enhancement algorithms, and then do subjective assessment in a pair-wise way to get the relative ranking of these enhanced images. A rank function is trained to fit the subjective assessment results, and can be used to predict ranks of new enhanced images which indicate the relative quality of enhancement algorithms. The experimental results show that our proposed approach statistically outperforms state-of-the-art general-purpose NR-IQA algorithms.
****

Shadow Removal from Single RGB-D Images
Yao Xiao, Efstratios Tsougenis, Chi-Keung Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3011-3018
We present the first automatic method to remove shadows from single RGB-D images. Using normal cues directly derived from depth, we can remove hard and soft shadows while preserving surface texture and shading. Our key assumption is: pixels with similar normals, spatial locations and chromaticity should have similar colors. A modified nonlocal matching is used to compute a shadow confidence map that localizes well hard shadow boundary, thus handling hard and soft shadows within the same framework. We compare our results produced using state-of-the-art shadow removal on single RGB images, and intrinsic image decomposition on standard RGB-D datasets.
****

Manifold Based Dynamic Texture Synthesis from Extremely Few Samples
Hongteng Xu, Hongyuan Zha, Mark A. Davenport; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3019-3026
In this paper, we present a novel method to synthesize dynamic texture sequences from extremely few samples, e.g., merely two possibly disparate frames, leveraging both Markov Random Fields (MRFs) and manifold learning. Decomposing a textural image as a set of patches, we achieve dynamic texture synthesis by estimating sequences of temporal patches. We select candidates for each temporal patch from spatial patches based on MRFs and regard them as samples from a low-dimensional manifold. After mapping candidates to a low-dimensional latent space, we estimate the sequence of temporal patches by finding an optimal trajectory in the latent space. Guided by some key properties of trajectories of realistic temporal patches, we derive a curvature-based trajectory selection algorithm. In contrast to the methods based on MRFs or dynamic systems that rely on a large amount of samples, our method is able to deal with the case of extremely few samples and requires no training phase. We compare our method with the state of the art and show that our method not only exhibits superior performance on synthesizing textures but it also produces results with pleasing visual effects.
****

The Synthesizability of Texture Examples
Dengxin Dai, Hayko Riemenschneider, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3027-3034
Example-based texture synthesis (ETS) has been widely used to generate high quality textures of desired sizes from a small example. However, not all textures are equally well reproducible that way. We predict how synthesizable a particular

texture is by ETS. We introduce a dataset (21,302 textures) of which all images have been annotated in terms of their synthesizability. We design a set of texture features, such as 'textureness', homogeneity, repetitiveness, and irregularity, and train a predictor using these features on the data collection. This work is the first attempt to quantify this image property, and we find that texture synthesizability can be learned and predicted. We use this insight to trim images to parts that are more synthesizable. Also we suggest which texture synthesis method is best suited to synthesise a given texture. Our approach can be seen as 'winner-uses-all': picking one method among several alternatives, ending up with an overall superior ETS method. Such strategy could also be considered for other vision tasks: rather than building an even stronger method, choose from existing methods based on some simple preprocessing.

****

Reconstructing Evolving Tree Structures in Time Lapse Sequences

Przemyslaw Glowacki, Miguel Amavel Pinheiro, Engin Turetken, Raphael Sznitman, Daniel Lebrecht, Jan Kybic, Anthony Holtmaat, Pascal Fua; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3035-3042

We propose an approach to reconstructing tree structures that evolve over time in 2D images and 3D image stacks such as neuronal axons or plant branches. Instead of reconstructing structures in each image independently, we do so for all images simultaneously to take advantage of temporal-consistency constraints. We show that this problem can be formulated as a Quadratic Mixed Integer Program and solved efficiently. The outcome of our approach is a framework that provides substantial improvements in reconstructions over traditional single time-instance formulations. Furthermore, an added benefit of our approach is the ability to automatically detect places where significant changes have occurred over time, which is challenging when considering large amounts of data.

****

Total-Variation Minimization on Unstructured Volumetric Mesh: Biophysical Applications on Reconstruction of 3D Ischemic Myocardium

Jingjia Xu, Azar Rahimi Dehaghani, Fei Gao, Linwei Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3043-3050

This paper describes the development and application of a new approach to total-variation (TV) minimization for reconstruction problems on geometrically-complex and unstructured volumetric mesh. The driving application of this study is the reconstruction of 3D ischemic regions in the heart from noninvasive body-surface potential data, where the use of a TV-prior can be expected to promote the reconstruction of two piecewise smooth regions of healthy and ischemic electrical properties with localized gradient in between. Compared to TV minimization on regular grids of pixels/voxels, the complex unstructured volumetric mesh of the heart poses unique challenges including the impact of mesh resolutions on the TV-prior and the difficulty of gradient calculation. In this paper, we introduce a variational TV-prior and, when combined with the iteratively re-weighted least-square concept, a new algorithm to TV minimization that is computationally efficient and robust to the discretization resolution. In a large set of simulation studies as well as two initial real-data studies, we show that the use of the proposed TV prior outperforms L2-based penalties in reconstruct ischemic regions, and it shows higher robustness and efficiency compared to the commonly used discrete TV prior. We also investigate the performance of the proposed TV-prior in combination with a L2- versus L1-based data fidelity term. The proposed method can extend TV-minimization to a border range of applications that involves physical domains of complex shape and unstructured volumetric mesh.

****

Tracking on the Product Manifold of Shape and Orientation for Tractography from Diffusion MRI

Yuanxiang Wang, Hesamoddin Salehian, Guang Cheng, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3051-3056

Tractography refers to the process of tracing out the nerve fiber bundles from diffusion Magnetic Resonance Images (dMRI) data acquired either in vivo or ex-vivo. Tractography is a mature research topic within the field of diffusion MRI analysis, nevertheless, several new methods are being proposed on a regular basis thereby justifying the need, as the problem is not fully solved. Tractography is usually applied to the model (used to represent the diffusion MR signal or a derived quantity) reconstructed from the acquired data. Separating shape and orientation of these models was previously shown to approximately preserve diffusion anisotropy (a useful bio-marker) in the ubiquitous problem of interpolation. However, no further intrinsic geometric properties of this framework were exploited to date in literature. In this paper, we propose a new intrinsic recursive filter on the product manifold of shape and orientation. The recursive filter, dubbed IUKFPro, is a generalization of the unscented Kalman filter (UKF) to this product manifold. The salient contributions of this work are: (1) A new intrinsic UKF for the product manifold of shape and orientation. (2) Derivation of the Riemannian geometry of the product manifold. (3) IUKFPro is tested on synthetic and real data sets from various tractography challenge competitions. From the experimental results, it is evident that IUKFPro performs better than several competing schemes in literature with regards to some of the error measures used in the competitions and is competitive with respect to others.

****

Curvilinear Structure Tracking by Low Rank Tensor Approximation with Model Propagation

Erkang Cheng, Yu Pang, Ying Zhu, Jingyi Yu, Haibin Ling; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3057-3064

Robust tracking of deformable object like catheter or vascular structures in X-ray images is an important technique used in image guided medical interventions for effective motion compensation and dynamic multi-modality image fusion. Tracking of such anatomical structures and devices is very challenging due to large degrees of appearance changes, low visibility of X-ray images and the deformable nature of the underlying motion field as a result of complex 3D anatomical movements projected into 2D images. To address these issues, we propose a new deformable tracking method using the tensor-based algorithm with model propagation. Specifically, the deformable tracking is formulated as a multi-dimensional assignment problem which is solved by rank-1 l1 tensor approximation. The model prior is propagated in the course of deformable tracking. Both the higher order information and the model prior provide powerful discriminative cues for reducing ambiguity arising from the complex background, and consequently improve the tracking robustness. To validate the proposed approach, we applied it to catheter and vascular structures tracking and tested on X-ray fluoroscopic sequences obtained from 17 clinical cases. The results show, both quantitatively and qualitatively, that our approach achieves a mean tracking error of 1.4 pixels for vascular structure and 1.3 pixels for catheter tracking.

****

Patch-based Evaluation of Image Segmentation

Christian Ledig, Wenzhe Shi, Wenjia Bai, Daniel Rueckert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3065-3072

The quantification of similarity between image segmentations is a complex yet important task. The ideal similarity measure should be unbiased to segmentations of different volume and complexity, and be able to quantify and visualise segmentation bias. Similarity measures based on overlap, e.g. Dice score, or surface distances, e.g. Hausdorff distance, clearly do not satisfy all of these properties. To address this problem, we introduce Patch-based Evaluation of Image Segmentation (PEIS), a general method to assess segmentation quality. Our method is based on finding patch correspondences and the associated patch displacements, which allow the estimation of segmentation bias. We quantify both the agreement of the segmentation boundary and the conservation of the segmentation shape. We further assess the segmentation complexity within patches to weight the contribution of local segmentation similarity to the global score. We evaluate PEIS on both synthetic data and two medical imaging datasets. On synthetic segmentations of different shapes, we provide evidence that PEIS, in comparison to the Dice score, produces more comparable scores, has increased sensitivity and estimates segmentation bias accurately. On cardiac magnetic resonance (MR) images, we demonstrate that PEIS can evaluate the performance of a segmentation method independent of the size or complexity of the segmentation under consideration. On brain MR images, we compare five different automatic hippocampus segmentation techniques using PEIS. Finally, we visualise the segmentation bias on a selection of the cases.

****

Evaluation of Scan-Line Optimization for 3D Medical Image Registration

Simon Hermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3073-3080

Scan-line optimization via cost accumulation has become very popular for stereo estimation in computer vision applications and is often combined with a semi-global cost integration strategy, known as SGM. This paper introduces this combination as a general and effective optimization technique. It is the first time that this concept is applied to 3D medical image registration. The presented algorithm, SGM-3D, employs a coarse-to-fine strategy and reduces the search space dimension for consecutive pyramid levels by a fixed linear rate. This allows it to handle large displacements to an extent that is required for clinical applications in high dimensional data. SGM-3D is evaluated in context of pulmonary motion analysis on the recently extended DIR-lab benchmark that provides ten 4D computed tomography (CT) image data sets, as well as ten challenging 3D CT scan pairs from the COPDgene study archive. Results show that both registration errors as well as run-time performance are very competitive with current state-of-the-art methods.

****

Classification of Histology Sections via Multispectral Convolutional Sparse Coding

Yin Zhou, Hang Chang, Kenneth Barner, Paul Spellman, Bahram Parvin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3081-3088

Image-based classification of histology sections plays an important role in predicting clinical outcomes. However this task is very challenging due to the presence of large technical variations (e.g., fixation, staining) and biological heterogeneities (e.g., cell type, cell state). In the field of biomedical imaging, for the purposes of visualization and/or quantification, different stains are typically used for different targets of interest (e.g., cellular/subcellular events), which generates multi-spectrum data (images) through various types of microscopes and, as a result, provides the possibility of learning biological-component-specific features by exploiting multispectral information. We propose a multispectral feature learning model that automatically learns a set of convolution filter banks from separate spectra to efficiently discover the intrinsic tissue morphometric signatures, based on convolutional sparse coding (CSC). The learned feature representations are then aggregated through the spatial pyramid matching framework (SPM) and finally classified using a linear SVM. The proposed system has been evaluated using two large-scale tumor cohorts, collected from The Cancer Genome Atlas (TCGA). Experimental results show that the proposed model 1) outperforms systems utilizing sparse coding for unsupervised feature learning (e.g., PSDSPM [5]); 2) is competitive with systems built upon features with biological prior knowledge (e.g., SMLSPM [4]).

****

Matrix-Similarity Based Loss Function and Feature Selection for Alzheimer's Disease Diagnosis

Xiaofeng Zhu, Heung-Il Suk, Dinggang Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3089-3096

Recent studies on Alzheimer's Disease (AD) or its prodromal stage, Mild Cognitive Impairment (MCI), diagnosis presented that the tasks of identifying brain disease status and predicting clinical scores based on neuroimaging features were h ighly related to each other. However, these tasks were often conducted independently in the previous studies. Regarding the feature selection, to our best knowledge, most of the previous work considered a loss function defined as an element-wise difference between the target values and the predicted ones. In this paper, we consider the problems of joint regression and classification for AD/MCI diagnosis and propose a novel matrix-similarity based loss function that uses high-level information inherent in the target response matrix and imposes the information to be preserved in the predicted response matrix. The newly devised loss function is combined with a group lasso method for joint feature selection across tasks, i.e., clinical scores prediction and disease status identification. We conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and showed that the newly devised loss function was effective to enhance the performances of both clinical score prediction and disease status identification, outperforming the state-of-the-art methods.

****

Discriminative Sparse Inverse Covariance Matrix: Application in Brain Functional Network Classification
Luping Zhou, Lei Wang, Philip Ogunbona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3097-3104
Recent studies show that mental disorders change the functional organization of the brain, which could be investigated via various imaging techniques. Analyzing such changes is becoming critical as it could provide new biomarkers for diagnosing and monitoring the progression of the diseases. Functional connectivity analysis studies the covary activity of neuronal populations in different brain regions. The sparse inverse covariance estimation (SICE), also known as graphical LASSO, is one of the most important tools for functional connectivity analysis, which estimates the interregional partial correlations of the brain. Although being increasingly used for predicting mental disorders, SICE is basically a generative method that may not necessarily perform well on classifying neuroimaging data. In this paper, we propose a learning framework to effectively improve the discriminative power of SICEs by taking advantage of the samples in the opposite class. We formulate our objective as convex optimization problems for both one-class and two-class classifications. By analyzing these optimization problems, we not only solve them efficiently in their dual form, but also gain insights into this new learning framework. The proposed framework is applied to analyzing the brain metabolic covariant networks built upon FDG-PET images for the prediction of the Alzheimer's disease, and shows significant improvement of classification performance for both one-class and two-class scenarios. Moreover, as SICE is a general method for learning undirected Gaussian graphical models, this paper has broader meanings beyond the scope of brain research.

****

A Bayesian Framework For the Local Configuration of Retinal Junctions
Touseef Ahmad Qureshi, Andrew Hunter, Bashir Al-Diri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3105-3110
Retinal images contain forests of mutually intersecting and overlapping venous and arterial vascular trees. The geometry of these trees shows adaptation to vascular diseases including diabetes, stroke and hypertension. Segmentation of the retinal vascular network is complicated by inconsistent vessel contrast, fuzzy edges, variable image quality, media opacities, complex intersections and overlaps. This paper presents a Bayesian approach to resolving the configuration of vascular junctions to correctly construct the vascular trees. A probabilistic model of vascular joints (terminals, bridges and bifurcations) and their configuration in junctions is built, and Maximum A Posteriori (MAP) estimation used to select most likely configurations. The model is built using a reference set of 3010 joints extracted from the DRIVE public domain vascular segmentation dataset, and evaluated on 3435 joints from the DRIVE test set, demonstrating an accuracy of 95.2%.

****

Learning-Based Atlas Selection for Multiple-Atlas Segmentation
Gerard Sanroma, Guorong Wu, Yaozong Gao, Dinggang Shen; Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3111-3117

Recently, multi-atlas segmentation (MAS) has achieved a great success in the medical imaging area. The key assumption of MAS is that multiple atlases encompass richer anatomical variability than a single atlas. Therefore, we can label the target image more accurately by mapping the label information from the appropriate atlas images that have the most similar structures. The problem of atlas selection, however, still remains unexplored. Current state-of-the-art MAS methods rely on image similarity to select a set of atlases. Unfortunately, this heuristic criterion is not necessarily related to segmentation performance and, thus may undermine segmentation results. To solve this simple but critical problem, we propose a learning-based atlas selection method to pick up the best atlases that would eventually lead to more accurate image segmentation. Our idea is to learn the relationship between the pairwise appearance of observed instances (a pair of atlas and target images) and their final labeling performance (in terms of Dice ratio). In this way, we can select the best atlases according to their expected labeling accuracy. It is worth noting that our atlas selection method is general enough to be integrated with existing MAS methods. As is shown in the experiments, we achieve significant improvement after we integrate our method with 3 widely used MAS methods on ADNI and LONI LPBA40 datasets.

****

Fully Automated Non-rigid Segmentation with Distance Regularized Level Set Evolution Initialized and Constrained by Deep-structured Inference

Tuan Anh Ngo, Gustavo Carneiro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3118-3125

We propose a new fully automated non-rigid segmentation approach based on the distance regularized level set method that is initialized and constrained by the results of a structured inference using deep belief networks. This recently proposed level-set formulation achieves reasonably accurate results in several segmentation problems, and has the advantage of eliminating periodic re-initializations during the optimization process, and as a result it avoids numerical errors. Nevertheless, when applied to challenging problems, such as the left ventricle segmentation from short axis cine magnetic ressonance (MR) images, the accuracy obtained by this distance regularized level set is lower than the state of the art. The main reasons behind this lower accuracy are the dependence on good initial guess for the level set optimization and on reliable appearance models. We address these two issues with an innovative structured inference using deep belief networks that produces reliable initial guess and appearance model. The effectiveness of our method is demonstrated on the MICCAI 2009 left ventricle segmentation challenge, where we show that our approach achieves one of the most competitive results (in terms of segmentation accuracy) in the field.

****

FAST LABEL: Easy and Efficient Solution of Joint Multi-Label and Estimation Problems

Ganesh Sundaramoorthi, Byung-Woo Hong; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3126-3133

We derive an easy-to-implement and efficient algorithm for solving multi-label image partitioning problems in the form of the problem addressed by Region Competition. These problems jointly determine a parameter for each of the regions in the partition. Given an estimate of the parameters, a fast approximate solution to the multi-label sub-problem is derived by a global update that uses smoothing and thresholding. The method is empirically validated to be robust to fine details of the image that plague local solutions. Further, in comparison to global methods for the multi-label problem, the method is more efficient and it is easy for a non-specialist to implement. We give sample Matlab code for the multi-label Chan-Vese problem in this paper! Experimental comparison to the state-of-the-art in multi-label solutions to Region Competition shows that our method achieves equal or better accuracy, with the main advantage being speed and ease of implementation.

****

Learning to Group Objects

Victoria Yanulevskaya, Jasper Uijlings, Nicu Sebe; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3134-3141

This paper presents a novel method to generate a hypothesis set of class-independent object regions. It has been shown that such object regions can be used to focus computer vision techniques on the parts of an image that matter most leading to significant improvements in both object localisation and semantic segmentation in recent years. Of course, the higher quality of class-independent object regions, the better subsequent computer vision algorithms can perform. In this paper we focus on generating higher quality object hypotheses. We start from an oversegmentation for which we propose to extract a wide variety of region-features. We group regions together in a hierarchical fashion, for which we train a Random Forest which predicts at each stage of the hierarchy the best possible merge. Hence unlike other approaches, we use relatively powerful features and classifiers at an early stage of the generation of likely object regions. Finally, we identify and combine stable regions in order to capture objects which consist of dissimilar parts. We show on the PASCAL 2007 and 2012 datasets that our method yields higher quality regions than competing approaches while it is at the same time more computationally efficient.

****

Unsupervised Multi-Class Joint Image Segmentation

Fan Wang, Qixing Huang, Maks Ovsjanikov, Leonidas J. Guibas; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3142-3149

Joint segmentation of image sets is a challenging problem, especially when there are multiple objects with variable appearance shared among the images in the collection and the set of objects present in each particular image is itself varying and unknown. In this paper, we present a novel method to jointly segment a set of images containing objects from multiple classes. We first establish consistent functional maps across the input images, and introduce a formulation that explicitly models partial similarity across images instead of global consistency. Given the optimized maps between pairs of images, multiple groups of consistent segmentation functions are found such that they align with segmentation cues in the images, agree with the functional maps, and are mutually exclusive. The proposed fully unsupervised approach exhibits a significant improvement over the state-of-the-art methods, as shown on the co-segmentation data sets MSRC, Flickr, and PASCAL.

****

Semantic Object Selection

Ejaz Ahmed, Scott Cohen, Brian Price; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3150-3157

Interactive object segmentation has great practical importance in computer vision. Many interactive methods have been proposed utilizing user input in the form of mouse clicks and mouse strokes, and often requiring a lot of user intervention. In this paper, we present a system with a far simpler input method: the user needs only give the name of the desired object. With the tag provided by the user we do a text query of an image database to gather exemplars of the object. Using object proposals and borrowing ideas from image retrieval and object detection, the object is localized in the target image. An appearance model generated from the exemplars and the location prior are used in an energy minimization framework to select the object. Our method outperforms the state-of-the-art on existing datasets and on a more challenging dataset we collected.

****

Discrete-Continuous Gradient Orientation Estimation for Faster Image Segmentation

Michael Donoser, Dieter Schmalstieg; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3158-3165

The state-of-the-art in image segmentation builds hierarchical segmentation structures based on analyzing local feature cues in spectral settings. Due to their impressive performance, such segmentation approaches have become building blocks in many computer vision applications. Nevertheless, the main bottlenecks are still the computationally demanding processes of local feature processing and spectral analysis. In this paper, we demonstrate that based on a discrete-continuous optimization of oriented gradient signals, we are able to provide segmentation performance competitive to state-of-the-art on BSDS 500 (even without any spectral analysis) while reducing computation time by a factor of 40 and memory demands by a factor of 10.

****

Object-based Multiple Foreground Video Co-segmentation
Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3166-3173

We present a video co-segmentation method that uses category-independent object proposals as its basic element and can extract multiple foreground objects in a video set. The use of object elements overcomes limitations of low-level feature representations in separating complex foregrounds and backgrounds. We formulate object-based co-segmentation as a co-selection graph in which regions with foreground-like characteristics are favored while also accounting for intra-video and inter-video foreground coherence. To handle multiple foreground objects, we expand the co-selection graph model into a proposed multi-state selection graph model (MSG) that optimizes the segmentations of different objects jointly. This extension into the MSG can be applied not only to our co-selection graph, but also can be used to turn any standard graph model into a multi-state selection solution that can be optimized directly by the existing energy minimization techniques. Our experiments show that our object-based multiple foreground video co-segmentation method (ObMiC) compares well to related techniques on both single and multiple foreground cases.

****

Parsing World's Skylines using Shape-Constrained MRFs
Rashmi Tonge, Subhransu Maji, C. V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3174-3181

We propose an approach for segmenting the individual buildings in typical skyline images. Our approach is based on a Markov Random Field (MRF) formulation that exploits the fact that such images contain overlapping objects of similar shapes exhibiting a "tiered" structure. Our contributions are the following: (1) A dataset of 120 high-resolution skyline images from twelve different cities with over 4,000 individually labeled buildings that allows us to quantitatively evaluate the performance of various segmentation methods, (2) An analysis of low-level features that are useful for segmentation of buildings, and (3) A shape-constrained MRF formulation that enforces shape priors over the regions. For simple shapes such as rectangles, our formulation is significantly faster to optimize than a standard MRF approach, while also being more accurate. We experimentally evaluate various MRF formulations and demonstrate the effectiveness of our approach in segmenting skyline images.

****

Clothing Co-Parsing by Joint Image Segmentation and Labeling
Wei Yang, Ping Luo, Liang Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3182-3189

This paper aims at developing an integrated system of clothing co-parsing, in order to jointly parse a set of clothing images (unsegmented but annotated with tags) into semantic configurations. We propose a data-driven framework consisting of two phases of inference. The first phase, referred as "image co-segmentation", iterates to extract consistent regions on images and jointly refines the regions over all images by employing the exemplar-SVM (ESVM) technique [23]. In the second phase (i.e. "region colabeling"), we construct a multi-image graphical model by taking the segmented regions as vertices, and incorporate several contexts of clothing configuration (e.g., item location and mutual interactions). The joint label assignment can be solved using the efficient Graph Cuts algorithm. In addition to evaluate our framework on the Fashionista dataset [30], we construct a dataset called CCP consisting of 2098 high-resolution street fashion photos to demonstrate the performance of our system. We achieve 90.29% / 88.23% segmenta tion accuracy and 65.52% / 63.89% recognition rate on the Fashionista and the CCP datasets, respectively, which are superior compared with state-of-the-art methods.

****

Tell Me What You See and I will Show You Where It Is

Jia Xu, Alexander G. Schwing, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3190-3197

We tackle the problem of weakly labeled semantic segmentation, where the only source of annotation are image tags encoding which classes are present in the scene. This is an extremely difficult problem as no pixel-wise labelings are available, not even at training time. In this paper, we show that this problem can be formalized as an instance of learning in a latent structured prediction framework, where the graphical model encodes the presence and absence of a class as well as the assignments of semantic labels to superpixels. As a consequence, we are able to leverage standard algorithms with good theoretical properties. We demonstrate the effectiveness of our approach using the challenging SIFT-flow dataset and show average per-class accuracy improvements of 7% over the state-of-the-art.

****

Beat the MTurkers: Automatic Image Labeling from Weak 3D Supervision

Liang-Chieh Chen, Sanja Fidler, Alan L. Yuille, Raquel Urtasun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3198-3205

Labeling large-scale datasets with very accurate object segmentations is an elaborate task that requires a high degree of quality control and a budget of tens or hundreds of thousands of dollars. Thus, developing solutions that can automatically perform the labeling given only weak supervision is key to reduce this cost. In this paper, we show how to exploit 3D information to automatically generate very accurate object segmentations given annotated 3D bounding boxes. We formulate the problem as the one of inference in a binary Markov random field which exploits appearance models, stereo and/or noisy point clouds, a repository of 3D CAD models as well as topological constraints. We demonstrate the effectiveness of our approach in the context of autonomous driving, and show that we can segment cars with the accuracy of 86% intersection-over-union, performing as well as highly recommended MTurkers!

****

Efficient Structured Parsing of Facades Using Dynamic Programming

Andrea Cohen, Alexander G. Schwing, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3206-3213

We propose a sequential optimization technique for segmenting a rectified image of a facade into semantic categories. Our method retrieves a parsing which respects common architectural constraints and also returns a certificate for global optimality. Contrasting the suggested method, the considered facade labeling problem is typically tackled as a classification task or as grammar parsing. Both approaches are not capable of fully exploiting the regularity of the problem. Therefore, our technique very significantly improves the accuracy compared to the state-of-the-art while being an order of magnitude faster. In addition, in 85% of the test images we obtain a certificate for optimality.

****

Dense Semantic Image Segmentation with Objects and Attributes

Shuai Zheng, Ming-Ming Cheng, Jonathan Warrell, Paul Sturgess, Vibhav Vineet, Carsten Rother, Philip H. S. Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3214-3221

The concepts of objects and attributes are both important for describing images precisely, since verbal descriptions often contain both adjectives and nouns (e.g. "I see a shiny red chair'). In this paper, we formulate the problem of joint visual attribute and object class image segmentation as a dense multi-labelling problem, where each pixel in an image can be associated with both an object-class and a set of visual attributes labels. In order to learn the label correlations, we adopt a boosting-based piecewise training approach with respect to the vis

ual appearance and co-occurrence cues. We use a filtering-based mean-field approximation approach for efficient joint inference. Further, we develop a hierarchical model to incorporate region-level object and attribute information. Experiments on the aPASCAL, CORE and attribute augmented NYU indoor scenes datasets show that the proposed approach is able to achieve state-of-the-art results.
****

Diffuse Mirrors: 3D Reconstruction from Diffuse Indirect Illumination Using Inexpensive Time-of-Flight Sensors

Felix Heide, Lei Xiao, Wolfgang Heidrich, Matthias B. Hullin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3222-3229

The functional difference between a diffuse wall and a mirror is well understood: one scatters back into all directions, and the other one preserves the directionality of reflected light. The temporal structure of the light, however, is left intact by both: assuming simple surface reflection, photons that arrive first are reflected first. In this paper, we exploit this insight to recover objects outside the line of sight from second-order diffuse reflections, effectively turning walls into mirrors. We formulate the reconstruction task as a linear inverse problem on the transient response of a scene, which we acquire using an affordable setup consisting of a modulated light source and a time-of-flight image sensor. By exploiting sparsity in the reconstruction domain, we achieve resolutions in the order of a few centimeters for object shape (depth and laterally) and albedo. Our method is robust to ambient light and works for large room-sized scenes. It is drastically faster and less expensive than previous approaches using femtosecond lasers and streak cameras, and does not require any moving parts.
****

Fourier Analysis on Transient Imaging with a Multifrequency Time-of-Flight Camera

Jingyu Lin, Yebin Liu, Matthias B. Hullin, Qionghai Dai; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3230-3237

A transient image is the optical impulse response of a scene which visualizes light propagation during an ultra-short time interval. In this paper we discover that the data captured by a multifrequency time-of-flight (ToF) camera is the Fourier transform of a transient image, and identify the sources of systematic error. Based on the discovery we propose a novel framework of frequency-domain transient imaging, as well as algorithms to remove systematic error. The whole process of our approach is of much lower computational cost, especially lower memory usage, than Heide et al.'s approach using the same device. We evaluate our approach on both synthetic and real-datasets.
****

Transparent Object Reconstruction via Coded Transport of Intensity

Chenguang Ma, Xing Lin, Jinli Suo, Qionghai Dai, Gordon Wetzstein; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3238-3245

Capturing and understanding visual signals is one of the core interests of computer vision. Much progress has been made w.r.t. many aspects of imaging, but the reconstruction of refractive phenomena, such as turbulence, gas and heat flows, liquids, or transparent solids, has remained a challenging problem. In this paper, we derive an intuitive formulation of light transport in refractive media using light fields and the transport of intensity equation. We show how coded illumination in combination with pairs of recorded images allow for robust computational reconstruction of dynamic two and three-dimensional refractive phenomena.
****

3D Shape and Indirect Appearance by Structured Light Transport

Matthew O'Toole, John Mather, Kiriakos N. Kutulakos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3246-3253

We consider the problem of deliberately manipulating the direct and indirect light flowing through a time-varying, fully-general scene in order to simplify its visual analysis. Our approach rests on a crucial link between stereo geometry an d light transport: while direct light always obeys the epipolar geometry of a projector-camera pair, indirect light overwhelmingly does not. We show that it is possible to turn this observation into an imaging method that analyzes light transport in real time in the optical domain, prior to acquisition. This yields three key abilities that we demonstrate in an experimental camera prototype: (1) producing a live indirect-only video stream for any scene, regardless of geometric or photometric complexity; (2) capturing images that make existing structured-light shape recovery algorithms robust to indirect transport; and (3) turning them into one-shot methods for dynamic 3D shape capture.

****

Shape-Preserving Half-Projective Warps for Image Stitching
Che-Han Chang, Yoichi Sato, Yung-Yu Chuang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3254-3261

This paper proposes a novel parametric warp which is a spatial combination of a projective transformation and a similarity transformation. Given the projective transformation relating two input images, based on an analysis of the projective transformation, our method smoothly extrapolates the projective transformation of the overlapping regions into the non-overlapping regions and the resultant warp gradually changes from projective to similarity across the image. The proposed warp has the strengths of both projective and similarity warps. It provides good alignment accuracy as projective warps while preserving the perspective of individual image as similarity warps. It can also be combined with more advanced local-warp-based alignment methods such as the as-projective-as-possible warp for better alignment accuracy. With the proposed warp, the field of view can be extended by stitching images with less projective distortion (stretched shapes and enlarged sizes).

****

Parallax-tolerant Image Stitching
Fan Zhang, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3262-3269

Parallax handling is a challenging task for image stitching. This paper presents a local stitching method to handle parallax based on the observation that input images do not need to be perfectly aligned over the whole overlapping region for stitching. Instead, they only need to be aligned in a way that there exists a local region where they can be seamlessly blended together. We adopt a hybrid alignment model that combines homography and content-preserving warping to provide flexibility for handling parallax and avoiding objectionable local distortion. We then develop an efficient randomized algorithm to search for a homography, which, combined with content-preserving warping, allows for optimal stitching. We predict how well a homography enables plausible stitching by finding a plausible seam and using the seam cost as the quality metric. We develop a seam finding method that estimates a plausible seam from only roughly aligned images by considering both geometric alignment and image content. We then pre-align input images using the optimal homography and further use content-preserving warping to locally refine the alignment. We finally compose aligned images together using a standard seam-cutting algorithm and a multi-band blending algorithm. Our experiments show that our method can effectively stitch images with large parallax that are difficult for existing methods.

****

Learning Everything about Anything: Webly-Supervised Visual Concept Learning
Santosh K. Divvala, Ali Farhadi, Carlos Guestrin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3270-3277

Recognition is graduating from labs to real-world applications. While it is encouraging to see its potential being tapped, it brings forth a fundamental challenge to the vision researcher: scalability. How can we learn a model for any concept that exhaustively covers all its appearance variations, while requiring minimal or no human supervision for compiling the vocabulary of visual variance, gathering the training images and annotations, and learning the models? In this paper, we introduce a fully-automated approach for learning extensive models for a wide range of variations (e.g. actions, interactions, attributes and beyond) with in any concept. Our approach leverages vast resources of online books to discover the vocabulary of variance, and intertwines the data collection and modeling steps to alleviate the need for explicit human supervision in training the models. Our approach organizes the visual knowledge about a concept in a convenient and useful way, enabling a variety of applications across vision and NLP. Our online system has been queried by users to learn models for several interesting concepts including breakfast, Gandhi, beautiful, etc. To date, our system has models available for over 50,000 variations within 150 concepts, and has annotated more than 10 million images with bounding boxes.

****

Dirichlet-based Histogram Feature Transform for Image Classification
Takumi Kobayashi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3278-3285

Histogram-based features have significantly contributed to recent development of image classifications, such as by SIFT local descriptors. In this paper, we propose a method to efficiently transform those histogram features for improving the classification performance. The (L1-normalized) histogram feature is regarded as a probability mass function, which is modeled by Dirichlet distribution. Based on the probabilistic modeling, we induce the Dirichlet Fisher kernel for transforming the histogram feature vector. The method works on the individual histogram feature to enhance the discriminative power at a low computational cost. On the other hand, in the bag-of-feature (BoF) framework, the Dirichlet mixture model can be extended to Gaussian mixture by transforming histogram-based local descriptors, e.g., SIFT, and thereby we propose the method of Dirichlet-derived GMM Fisher kernel. In the experiments on diverse image classification tasks including recognition of subordinate objects and material textures, the proposed methods improve the performance of the histogram-based features and BoF-based Fisher kernel, being favorably competitive with the state-of-the-arts.

****

BING: Binarized Normed Gradients for Objectness Estimation at 300fps
Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, Philip Torr; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3286-3293

Training a generic objectness measure to produce a small set of candidate object windows, has been shown to speed up the classical sliding window object detection paradigm. We observe that generic objects with well-defined closed boundary can be discriminated by looking at the norm of gradients, with a suitable resizing of their corresponding image windows in to a small fixed size. Based on this observation and computational reasons, we propose to resize the window to 8 Ã■ 8 and use the norm of the gradients as a simple 64D feature to describe it, for explicitly training a generic objectness measure. We further show how the binarized version of this feature, namely binarized normed gradients (BING), can be used for efficient objectness estimation, which requires only a few atomic operations (e.g. ADD, BITWISE SHIFT, etc.). Experiments on the challenging PASCAL VOC 2007 dataset show that our method efficiently (300fps on a single laptop CPU) generates a small set of category-independent, high quality object windows, yielding 96.2% object detection rate (DR) with 1,000 proposals. Increasing the numbers of proposals and color spaces for computing BING features, our performance can be further improved to 99.5% DR.

****

Context Driven Scene Parsing with Attention to Rare Classes
Jimei Yang, Brian Price, Scott Cohen, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3294-3301

This paper presents a scalable scene parsing algorithm based on image retrieval and superpixel matching. We focus on rare object classes, which play an important role in achieving richer semantic understanding of visual scenes, compared to common background classes. Towards this end, we make two novel contributions: rare class expansion and semantic context description. First, considering the long-tailed nature of the label distribution, we expand the retrieval set by rare class exemplars and thus achieve more balanced superpixel classification results.

Second, we incorporate both global and local semantic context information through a feedback based mechanism to refine image retrieval and superpixel matching. Results on the SIFTflow and LMSun datasets show the superior performance of our algorithm, especially on the rare classes, without sacrificing overall labeling accuracy.

****

Patch to the Future: Unsupervised Visual Prediction

Jacob Walker, Abhinav Gupta, Martial Hebert; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3302-3309

In this paper we present a conceptually simple but surprisingly powerful method for visual prediction which combines the effectiveness of mid-level visual elements with temporal modeling. Our framework can be learned in a completely unsupervised manner from a large collection of videos. However, more importantly, because our approach models the prediction framework on these mid-level elements, we can not only predict the possible motion in the scene but also predict visual appearances â■■ how are appearances going to change with time. This yields a visual "hallucination" of probable events on top of the scene. We show that our method is able to accurately predict and visualize simple future events; we also show that our approach is comparable to supervised methods for event prediction.

****

Triangulation Embedding and Democratic Aggregation for Image Search

Herve Jegou, Andrew Zisserman; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3310-3317

We consider the design of a single vector representation for an image that embeds and aggregates a set of local patch descriptors such as SIFT. More specifically we aim to construct a dense representation, like the Fisher Vector or VLAD, though of small or intermediate size. We make two contributions, both aimed at regularizing the individual contributions of the local descriptors in the final representation. The first is a novel embedding method that avoids the dependency on absolute distances by encoding directions. The second contribution is a "democratization" strategy that further limits the interaction of unrelated descriptors in the aggregation stage. These methods are complementary and give a substantial performance boost over the state of the art in image search with short or mid-size vectors, as demonstrated by our experiments on standard public image retrieval benchmarks.

****

Low-Cost Compressive Sensing for Color Video and Depth

Xin Yuan, Patrick Llull, Xuejun Liao, Jianbo Yang, David J. Brady, Guillermo Sapiro, Lawrence Carin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3318-3325

A simple and inexpensive (low-power and low-bandwidth) modification is made to a conventional off-the-shelf color video camera, from which we recover multiple color frames for each of the original measured frames, and each of the recovered frames can be focused at a different depth. The recovery of multiple frames for each measured frame is made possible via high-speed coding, manifested via translation of a single coded aperture; the inexpensive translation is constituted by mounting the binary code on a piezoelectric device. To simultaneously recover depth information, a liquid lens is modulated at high speed, via a variable voltage. Consequently, during the aforementioned coding process, the liquid lens allows the camera to sweep the focus through multiple depths. In addition to designing and implementing the camera, fast recovery is achieved by an anytime algorithm exploiting the group-sparsity of wavelet/DCT coefficients.

****

Aliasing Detection and Reduction in Plenoptic Imaging

Zhaolin Xiao, Qing Wang, Guoqing Zhou, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3326-3333

When using plenoptic camera for digital refocusing, angular undersampling can cause severe (angular) aliasing artifacts. Previous approaches have focused on avoiding aliasing by pre-processing the acquired light field via prefiltering, demosaicing, reparameterization, etc. In this paper, we present a different solution that first detects and then removes aliasing at the light field refocusing stage. Different from previous frequency domain aliasing analysis, we carry out a spatial domain analysis to reveal whether the aliasing would occur and uncover where in the image it would occur. The spatial analysis also facilitates easy separation of the aliasing vs. non-aliasing regions and aliasing removal. Experiments on both synthetic scene and real light field camera array data sets demonstrate that our approach has a number of advantages over the classical prefiltering and depth-dependent light field rendering techniques.

****

Illumination-Aware Age Progression

Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, Steven M. Seitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3334-3341

We present an approach that takes a single photograph of a child as input and automatically produces a series of age-progressed outputs between 1 and 80 years of age, accounting for pose, expression, and illumination. Leveraging thousands of photos of children and adults at many ages from the Internet, we first show how to compute average image subspaces that are pixel-to-pixel aligned and model variable lighting. These averages depict a prototype man and woman aging from 0 to 80, under any desired illumination, and capture the differences in shape and texture between ages. Applying these differences to a new photo yields an age progressed result.   Contributions include relightable age subspaces, a novel technique for subspace-to-subspace alignment, and the most extensive evaluation of age progression techniques in the literature.

****

Color Transfer Using Probabilistic Moving Least Squares

Youngbae Hwang, Joon-Young Lee, In So Kweon, Seon Joo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3342-3349

This paper introduces a new color transfer method which is a process of transferring color of an image to match the color of another image of the same scene. The color of a scene may vary from image to image because the photographs are taken at different times, with different cameras, and under different camera settings. To solve for a full nonlinear and nonparametric color mapping in the 3D RGB color space, we propose a scattered point interpolation scheme using moving least squares and strengthen it with a probabilistic modeling of the color transfer in the 3D color space to deal with mis-alignments and noise. Experiments show the effectiveness of our method over previous color transfer methods both quantitatively and qualitatively. In addition, our framework can be applied for various instances of color transfer such as transferring color between different camera models, camera settings, and illumination conditions, as well as for video color transfers.

****

Image Pre-compensation: Balancing Contrast and Ringing

Yu Ji, Jinwei Ye, Sing Bing Kang, Jingyi Yu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3350-3357

The goal of image pre-compensation is to process an image such that after being convolved with a known kernel, will appear close to the sharp reference image. In a practical setting, the pre-compensated image has significantly higher dynamic range than the latent image. As a result, some form of tone mapping is needed. In this paper, we show how global tone mapping functions affect contrast and ringing in image pre-compensation. In particular, we show that linear tone mapping eliminates ringing but incurs severe contrast loss, while non-linear tone mapping functions such as Gamma curves slightly enhances contrast but introduces ringing. To enable quantitative analysis, we design new metrics to measure the contrast of an image with ringing. Specifically, we set out to find its "equivalent ringing-free" image that matches its intensity histogram and uses its contrast as the measure. We illustrate our approach on projector defocus compensation and visual acuity enhancement. Compared with the state-of-the-art, our approach significantly improves the contrast. We believe our technique is the first to analyti

cally trade-off between contrast and ringing.
****

Time-Mapping Using Space-Time Saliency
Feng Zhou, Sing Bing Kang, Michael F. Cohen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3358-3365

We describe a new approach for generating regular-speed, low-frame-rate (LFR) video from a high-frame-rate (HFR) input while preserving the important moments in the original. We call this time-mapping, a time-based analogy to high dynamic range to low dynamic range spatial tone-mapping. Our approach makes these contributions: (1) a robust space-time saliency method for evaluating visual importance, (2) a re-timing technique to temporally resample based on frame importance, and (3) temporal filters to enhance the rendering of salient motion. Results of our space-time saliency method on a benchmark dataset show it is state-of-the-art. In addition, the benefits of our approach to HFR-to-LFR time-mapping over more direct methods are demonstrated in a user study.
****

Gyro-Based Multi-Image Deconvolution for Removing Handshake Blur
Sung Hee Park, Marc Levoy; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3366-3373

Image deblurring to remove blur caused by camera shake has been intensively studied. Nevertheless, most methods are brittle and computationally expensive. In this paper we analyze multi-image approaches, which capture and combine multiple frames in order to make deblurring more robust and tractable. In particular, we compare the performance of two approaches: align-and-average and multi-image deconvolution. Our deconvolution is non-blind, using a blur model obtained from real camera motion as measured by a gyroscope. We show that in most situations such deconvolution outperforms align-and-average. We also show, perhaps surprisingly, that deconvolution does not benefit from increasing exposure time beyond a certain threshold. To demonstrate the effectiveness and efficiency of our method, we apply it to still-resolution imagery of natural scenes captured using a mobile camera with flexible camera control and an attached gyroscope.
****

Similarity-Aware Patchwork Assembly for Depth Image Super-Resolution
Jing Li, Zhichao Lu, Gang Zeng, Rui Gan, Hongbin Zha; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3374-3381

This paper describes a patchwork assembly algorithm for depth image super-resolution. An input low resolution depth image is disassembled into parts by matching similar regions on a set of high resolution training images, and a super-resolution image is then assembled using these corresponding matched counterparts. We convert the super resolution problem into a Markov Random Field (MRF) labeling problem, and propose a unified formulation embedding (1) the consistency between the resolution enhanced image and the original input, (2) the similarity of disassembled parts with the corresponding regions on training images, (3) the depth smoothness in local neighborhoods, (4) the additional geometric constraints from self-similar structures in the scene, and (5) the boundary coincidence between the resolution enhanced depth image and an optional aligned high resolution intensity image. Experimental results on both synthetic and real-world data demonstrate that the proposed algorithm is capable of recovering high quality depth images with X4 resolution enhancement along each coordinate direction, and that it outperforms state-of-the-arts [14] in both qualitative and quantitative evaluations.
****

Deblurring Low-light Images with Light Streaks
Zhe Hu, Sunghyun Cho, Jue Wang, Ming-Hsuan Yang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3382-3389

Images taken in low-light conditions with handheld cameras are often blurry due to the required long exposure time. Although significant progress has been made recently on image deblurring, state-of-the-art approaches often fail on low-light images, as these images do not contain a sufficient number of salient features that deblurring methods rely on. On the other hand, light streaks are common ph enomena in low-light images that contain rich blur information, but have not been extensively explored in previous approaches. In this work, we propose a new method that utilizes light streaks to help deblur low-light images. We introduce a non-linear blur model that explicitly models light streaks and their underlying light sources, and poses them as constraints for estimating the blur kernel in an optimization framework. Our method also automatically detects useful light streaks in the input image. Experimental results show that our approach obtains good results on challenging real-world examples that no other methods could achieve before.

****

Depth Enhancement via Low-rank Matrix Completion

Si Lu, Xiaofeng Ren, Feng Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3390-3397

Depth captured by consumer RGB-D cameras is often noisy and misses values at some pixels, especially around object boundaries. Most existing methods complete the missing depth values guided by the corresponding color image. When the color image is noisy or the correlation between color and depth is weak, the depth map cannot be properly enhanced. In this paper, we present a depth map enhancement algorithm that performs depth map completion and de-noising simultaneously. Our method is based on the observation that similar RGB-D patches lie in a very low-dimensional subspace. We can then assemble the similar patches into a matrix and enforce this low-rank subspace constraint. This low-rank subspace constraint essentially captures the underlying structure in the RGB-D patches and enables robust depth enhancement against the noise or weak correlation between color and depth. Based on this subspace constraint, our method formulates depth map enhancement as a low-rank matrix completion problem. Since the rank of a matrix changes over matrices, we develop a data-driven method to automatically determine the rank number for each matrix. The experiments on both public benchmarks and our own captured RGB-D images show that our method can effectively enhance depth maps.

****

Raw-to-Raw: Mapping between Image Sensor Color Responses

Rang Nguyen, Dilip K. Prasad, Michael S. Brown; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3398-3405

Camera images saved in raw format are being adopted in computer vision tasks since raw values represent minimally processed sensor responses. Camera manufacturers, however, have yet to adopt a standard for raw images and current raw-rgb values are device specific due to different sensors spectral sensitivities.   This results in significantly different raw images for the same scene captured with different cameras. This paper focuses on estimating a mapping that can convert a raw image of an arbitrary scene and illumination from one camera's raw space to another. To this end, we examine various mapping strategies including linear and non-linear transformations applied both in a global and illumination-specific manner. We show that illumination-specific mappings give the best result, however, at the expense of requiring a large number of transformations. To address this issue, we introduce an illumination-independent mapping approach that uses white-balancing to assist in reducing the number of required transformations. We show that this approach achieves state-of-the-art results on a range of consumer cameras and images of arbitrary scenes and illuminations.

****

DAISY Filter Flow: A Generalized Discrete Approach to Dense Correspondences

Hongsheng Yang, Wen-Yan Lin, Jiangbo Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3406-3413

Establishing dense correspondences reliably between a pair of images is an important vision task with many applications. Though significant advance has been made towards estimating dense stereo and optical flow fields for two images adjacent in viewpoint or in time, building reliable dense correspondence fields for two general images still remains largely unsolved. For instance, two given images sharing some content exhibit dramatic photometric and geometric variations, or they depict different 3D scenes of similar scene characteristics. Fundamental challenges to such an image or scene alignment task are often multifold, which rende r many existing techniques fall short of producing dense correspondences robustly and efficiently. This paper presents a novel approach called DAISY filter flow (DFF) to address this challenging task. Inspired by the recent PatchMatch Filter technique, we leverage and extend a few established methods: 1) DAISY descriptors, 2) filter-based efficient flow inference, and 3) the PatchMatch fast search. Coupling and optimizing these modules seamlessly with image segments as the bridge, the proposed DFF approach enables efficiently performing dense descriptor-based correspondence field estimation in a generalized high-dimensional label space, which is augmented by scales and rotations. Experiments on a variety of challenging scenes show that our DFF approach estimates spatially coherent yet discontinuity-preserving image alignment results both robustly and efficiently.

****

Robust 3D Tracking with Descriptor Fields

Alberto Crivellaro, Vincent Lepetit; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3414-3421

We introduce a method that can register challenging images from specular and poorly textured 3D environments, on which previous approaches fail. We assume that a small set of reference images of the environment and a partial 3D model are available. Like previous approaches, we register the input images by aligning them with one of the reference images using the 3D information. However, these approaches typically rely on the pixel intensities for the alignment, which is prone to fail in presence of specularities or in absence of texture. Our main contribution is an efficient novel local descriptor that we use to describe each image location. We show that we can rely on this descriptor in place of the intensities to significantly improve the alignment robustness at a minor increase of the computational cost, and we analyze the reasons behind the success of our descriptor.

****

Evolutionary Quasi-random Search for Hand Articulations Tracking

Iason Oikonomidis, Manolis I.A. Lourakis, Antonis A. Argyros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3422-3429

We present a new method for tracking the 3D position, global orientation and full articulation of human hands. Following recent advances in model-based, hypothesize-and-test methods, the high-dimensional parameter space of hand configurations is explored with a novel evolutionary optimization technique specifically tailored to the problem. The proposed method capitalizes on the fact that samples from quasi-random sequences such as the Sobol have low discrepancy and exhibit a more uniform coverage of the sampled space compared to random samples obtained from the uniform distribution. The method has been tested for the problems of tracking the articulation of a single hand (27D parameter space) and two hands (54D space). Extensive experiments have been carried out with synthetic and real data, in comparison with state of the art methods. The quantitative evaluation shows that for cases of limited computational resources, the new approach achieves a speed-up of four (single hand tracking) and eight (two hands tracking) without compromising tracking accuracy. Interestingly, the proposed method is preferable compared to the state of the art either in the case of limited computational resources or in the case of more complex (i.e., higher dimensional) problems, thus improving the applicability of the method in a number of application domains.

****

Scalable 3D Tracking of Multiple Interacting Objects

Nikolaos Kyriazis, Antonis Argyros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3430-3437

We consider the problem of tracking multiple interacting objects in 3D, using RGBD input and by considering a hypothesize-and-test approach. Due to their interaction, objects to be tracked are expected to occlude each other in the field of view of the camera observing them. A naive approach would be to employ a Set of Independent Trackers (SIT) and to assign one tracker to each object. This approach scales well with the number of objects but fails as occlusions become stronger due to their disjoint consideration. The solution representing the current sta te of the art employs a single Joint Tracker (JT) that accounts for all objects simultaneously. This directly resolves ambiguities due to occlusions but has a computational complexity that grows geometrically with the number of tracked objects. We propose a middle ground, namely an Ensemble of Collaborative Trackers (ECT), that combines best traits from both worlds to deliver a practical and accurate solution to the multi-object 3D tracking problem. We present quantitative and qualitative experiments with several synthetic and real world sequences of diverse complexity. Experiments demonstrate that ECT manages to track far more complex scenes than JT at a computational time that is only slightly larger than that of SIT.

****

Bayesian Active Appearance Models

Joan Alabort-i-Medina, Stefanos Zafeiriou; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3438-3445

In this paper we provide the first, to the best of our knowledge, Bayesian formulation of one of the most successful and well-studied statistical models of shape and texture, i.e. Active Appearance Models (AAMs). To this end, we use a simple probabilistic model for texture generation assuming both Gaussian noise and a Gaussian prior over a latent texture space. We retrieve the shape parameters by formulating a novel cost function obtained by marginalizing out the latent texture space. This results in a fast implementation when compared to other simultaneous algorithms for fitting AAMs, mainly due to the removal of the calculation of texture parameters. We demonstrate that, contrary to what is believed regarding the performance of AAMs in generic fitting scenarios, optimization of the proposed cost function produces results that outperform discriminatively trained state-of-the-art methods in the problem of facial alignment "in the wild".

****

Human Shape and Pose Tracking Using Keyframes

Chun-Hao Huang, Edmond Boyer, Nassir Navab, Slobodan Ilic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3446-3453

This paper considers human tracking in multi-view setups and investigates a robust strategy that learns online key poses to drive a shape tracking method. The interest arises in realistic dynamic scenes where occlusions or segmentation errors occur. The corrupted observations present missing data and outliers that deteriorate tracking results. We propose to use key poses of the tracked person as multiple reference models. In contrast to many existing approaches that rely on a single reference model, multiple templates represent a larger variability of human poses. They provide therefore better initial hypotheses when tracking with noisy data. Our approach identifies these reference models online as distinctive keyframes during tracking. The most suitable one is then chosen as the reference at each frame. In addition, taking advantage of the proximity between successive frames, an efficient outlier handling technique is proposed to prevent from associating the model to irrelevant outliers. The two strategies are successfully experimented with a surface deformation framework that recovers both the pose and the shape. Evaluations on existing datasets also demonstrate their benefits with respect to the state of the art.

****

Better Feature Tracking Through Subspace Constraints

Bryan Poling, Gilad Lerman, Arthur Szlam; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3454-3461

Feature tracking in video is a crucial task in computer vision. Usually, the tracking problem is handled one feature at a time, using a single-feature tracker like the Kanade-Lucas-Tomasi algorithm, or one of its derivatives. While this approach works quite well when dealing with high-quality video and "strong" features, it often falters when faced with dark and noisy video containing low-quality features. We present a framework for jointly tracking a set of features, which enables sharing information between the different features in the scene. We show that our method can be employed to track features for both rigid and non-rigid motions (possibly of few moving bodies) even when some features are occluded. Fur

thermore, it can be used to significantly improve tracking results in poorly-lit scenes (where there is a mix of good and bad features). Our approach does not require direct modeling of the structure or the motion of the scene, and runs in real time on a single CPU core.
****

Online Object Tracking, Learning and Parsing with And-Or Graphs
Yang Lu, Tianfu Wu, Song Chun Zhu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3462-3469

This paper presents a framework for simultaneously tracking, learning and parsing objects with a hierarchical and compositional And-Or graph (AOG) representation. The AOG is discriminatively learned online to account for the appearance (e.g., lighting and partial occlusion) and structural (e.g., different poses and viewpoints) variations of the object itself, as well as the distractors (e.g., similar objects) in the scene background. In tracking, the state of the object (i.e., bounding box) is inferred by parsing with the current AOG using a spatial-temporal dynamic programming (DP) algorithm. When the AOG grows big for handling objects with large variations in long-term tracking, we propose a bottom-up/top-down scheduling scheme for efficient inference, which performs focused inference with the most stable and discriminative small sub-AOG. During online learning, the AOG is re-learned iteratively with two steps: (i) Identifying the false positives and false negatives of the current AOG in a new frame by exploiting the spatial and temporal constraints observed in the trajectory; (ii) Updating the structure of the AOG, and re-estimating the parameters based on the augmented training dataset. In experiments, the proposed method outperforms state-of-theart tracking algorithms on a recent public tracking benchmark with 50 testing videos and 30 publicly available trackers evaluated [34].
****

Region-based Particle Filter for Video Object Segmentation
David Varas, Ferran Marques; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3470-3477

We present a video object segmentation approach that extends the particle filter to a region-based image representation. Image partition is considered part of the particle filter measurement, which enriches the available information and leads to a re-formulation of the particle filter. The prediction step uses a co-clustering between the previous image object partition and a partition of the current one, which allows us to tackle the evolution of non-rigid structures. Particles are defined as unions of regions in the current image partition and their propagation is computed through a single co-clustering. The proposed technique is assessed on the SegTrack dataset, leading to satisfactory perceptual results and obtaining very competitive pixel error rates compared with the state-of-the-art methods.
****

Visual Tracking via Probability Continuous Outlier Model
Dong Wang, Huchuan Lu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3478-3485

In this paper, we present a novel online visual tracking method based on linear representation. First, we present a novel probability continuous outlier model (PCOM) to depict the continuous outliers that occur in the linear representation model. In the proposed model, the element of the noisy observation sample can be either represented by a PCA subspace with small Guassian noise or treated as an arbitrary value with a uniform prior, in which the spatial consistency prior is exploited by using a binary Markov random field model. Then, we derive the objective function of the PCOM method, the solution of which can be iteratively obtained by the outlier-free least squares and standard max-flow/min-cut steps. Finally, based on the proposed PCOM method, we design an effective observation likelihood function and a simple update scheme for visual tracking. Both qualitative and quantitative evaluations demonstrate that our tracker achieves very favorable performance in terms of both accuracy and speed.
****

Visual Tracking Using Pertinent Patch Selection and Masking

Dae-Youn Lee, Jae-Young Sim, Chang-Su Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3486-3493

A novel visual tracking algorithm using patch-based appearance models is proposed in this paper. We first divide the bounding box of a target object into multiple patches and then select only pertinent patches, which occur repeatedly near the center of the bounding box, to construct the foreground appearance model. We also divide the input image into non-overlapping blocks, construct a background model at each block location, and integrate these background models for tracking. Using the appearance models, we obtain an accurate foreground probability map. Finally, we estimate the optimal object position by maximizing the likelihood, which is obtained by convolving the foreground probability map with the pertinence mask. Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art tracking algorithms significantly in terms of center position errors and success rates.

****

Interval Tracker: Tracking by Interval Analysis

Junseok Kwon, Kyoung Mu Lee; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3494-3501

This paper proposes a robust tracking method that uses interval analysis. Any single posterior model necessarily includes a modeling uncertainty (error), and thus, the posterior should be represented as an interval of probability. Then, the objective of visual tracking becomes to find the best state that maximizes the posterior and minimizes its interval simultaneously. By minimizing the interval of the posterior, our method can reduce the modeling uncertainty in the posterior. In this paper, the aforementioned objective is achieved by using the M4 estimation, which combines the Maximum a Posterior (MAP) estimation with Minimum Mean-Square Error (MMSE), Maximum Likelihood (ML), and Minimum Interval Length (MIL) estimations. In the M4 estimation, our method maximizes the posterior over the state obtained by the MMSE estimation. The method also minimizes interval of the posterior by reducing the gap between the lower and upper bounds of the posterior. The gap is reduced when the likelihood is maximized by the ML estimation and the interval length of the state is minimized by the MIL estimation. The experimental results demonstrate that M4 estimation can be easily integrated into conventional tracking methods and can greatly enhance their tracking accuracy. In several challenging datasets, our method outperforms state-of-the-art tracking methods.

****

Unifying Spatial and Attribute Selection for Distracter-Resilient Tracking

Nan Jiang, Ying Wu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3502-3509

Visual distracters are detrimental and generally very difficult to handle in target tracking, because they generate false positive candidates for target matching. The resilience of region-based matching to the distracters depends not only on the matching metric, but also on the characteristics of the target region to be matched. The two tasks, i.e., learning the best metric and selecting the distracter-resilient target regions, actually correspond to the attribute selection and spatial selection processes in the human visual perception. This paper presents an initial attempt to unify the modeling of these two tasks for an effective solution, based on the introduction of a new quantity called Soft Visual Margin. As a function of both matching metric and spatial location, it measures the discrimination between the target and its spatial distracters, and characterizes the reliability of matching. Different from other formulations of margin, this new quantity is analytical and is insensitive to noisy data. This paper presents a novel method to jointly determine the best spatial location and the optimal metric. Based on that, a solid distracter-resilient region tracker is designed, and its effectiveness is validated and demonstrated through extensive experiments.

****

Pedestrian Detection in Low-resolution Imagery by Learning Multi-scale Intrinsic Motion Structures (MIMS)

Jiejie Zhu, Omar Javed, Jingen Liu, Qian Yu, Hui Cheng, Harpreet Sawhney; Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3510-3517

Detecting pedestrians at a distance from large-format wide-area imagery is a challenging problem because of low ground sampling distance (GSD) and low frame rate of the imagery. In such a scenario, the approaches based on appearance cues alone mostly fail because pedestrians are only a few pixels in size. Frame-differencing and optical flow based approaches also give poor detection results due to noise, camera jitter and parallax in aerial videos. To overcome these challenges, we propose a novel approach to extract Multi-scale Intrinsic Motion Structure features from pedestrian's motion patterns for pedestrian detection. The MIMS feature encodes the intrinsic motion properties of an object, which are location, velocity and trajectory-shape invariant. The extracted MIMS representation is robust to noisy flow estimates. In this paper, we give a comparative evaluation of the proposed method and demonstrate that MIMS outperforms the state of the art approaches in identifying pedestrians from low resolution airborne videos.

****

Multi-target Tracking with Motion Context in Tensor Power Iteration

Xinchu Shi, Haibin Ling, Weiming Hu, Chunfeng Yuan, Junliang Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3518-3525

Interactions between moving targets often provide discriminative clues for multiple target tracking (MTT), though many existing approaches ignore such interactions due to difficulty in effectively handling them. In this paper, we model interactions between neighbor targets by pair-wise motion context, and further encode such context into the global association optimization. To solve the resulting global non-convex maximization, we propose an effective and efficient power iteration framework. This solution enjoys two advantages for MTT: First, it allows us to combine the global energy accumulated from individual trajectories and the between-trajectory interaction energy into a united optimization, which can be solved by the proposed power iteration algorithm. Second, the framework is flexible to accommodate various types of pairwise context models and we in fact studied two different context models in this paper. For evaluation, we apply the proposed methods to four public datasets involving different challenging scenarios such as dense aerial borne traffic tracking, dense point set tracking, and semi-crowded pedestrian tracking. In all the experiments, our approaches demonstrate very promising results in comparison with state-of-the-art trackers.

****

SphereFlow: 6 DoF Scene Flow from RGB-D Pairs

Michael Hornacek, Andrew Fitzgibbon, Carsten Rother; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3526-3533

We take a new approach to computing dense scene flow between a pair of consecutive RGB-D frames. We exploit the availability of depth data by seeking correspondences with respect to patches specified not as the pixels inside square windows, but as the 3D points that are the inliers of spheres in world space. Our primary contribution is to show that by reasoning in terms of such patches under 6 DoF rigid body motions in 3D, we succeed in obtaining compelling results at displacements large and small without relying on either of two simplifying assumptions that pervade much of the earlier literature: brightness constancy or local surface planarity. As a consequence of our approach, our output is a dense field of 3D rigid body motions, in contrast to the 3D translations that are the norm in scene flow. Reasoning in our manner additionally allows us to carry out occlusion handling using a 6 DoF consistency check for the flow computed in both directions and a patchwise silhouette check to help reason about alignments in occlusion areas, and to promote smoothness of the flow fields using an intuitive local rigidity prior. We carry out our optimization in two steps, obtaining a first correspondence field using an adaptation of PatchMatch, and subsequently using alpha-expansion to jointly handle occlusions and perform regularization. We show attractive flow results on challenging synthetic and real-world scenes that push the practical limits of the aforementioned assumptions.

****

Fast Edge-Preserving PatchMatch for Large Displacement Optical Flow

Linchao Bao, Qingxiong Yang, Hailin Jin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3534-3541

We present a fast optical flow algorithm that can handle large displacement motions. Our algorithm is inspired by recent successes of local methods in visual correspondence searching as well as approximate nearest neighbor field algorithms. The main novelty is a fast randomized edge-preserving approximate nearest neighbor field algorithm which propagates self-similarity patterns in addition to offsets. Experimental results on public optical flow benchmarks show that our method is significantly faster than state-of-the-art methods without compromising on quality, especially when scenes contain large motions.

****

Learning an Image-based Motion Context for Multiple People Tracking

Laura Leal-Taixe, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, Silvio Savarese; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3542-3549

We present a novel method for multiple people tracking that leverages a generalized model for capturing interactions among individuals. At the core of our model lies a learned dictionary of interaction feature strings which capture relationships between the motions of targets. These feature strings, created from low-level image features, lead to a much richer representation of the physical interactions between targets compared to hand-specified social force models that previous works have introduced for tracking. One disadvantage of using social forces is that all pedestrians must be detected in order for the forces to be applied, while our method is able to encode the effect of undetected targets, making the tracker more robust to partial occlusions. The interaction feature strings are used in a Random Forest framework to track targets according to the features surrounding them. Results on six publicly available sequences show that our method outperforms state-of-the-art approaches in multiple people tracking.

****

Semi-Supervised Coupled Dictionary Learning for Person Re-identification

Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, Jiajun Bu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3550-3557

The desirability of being able to search for specific persons in surveillance videos captured by different cameras has increasingly motivated interest in the problem of person re-identification, which is a critical yet under-addressed challenge in multi-camera tracking systems. The main difficulty of person re-identification arises from the variations in human appearances from different camera views. In this paper, to bridge the human appearance variations across cameras, two coupled dictionaries that relate to the gallery and probe cameras are jointly learned in the training phase from both labeled and unlabeled images. The labeled training images carry the relationship between features from different cameras, and the abundant unlabeled training images are introduced to exploit the geometry of the marginal distribution for obtaining robust sparse representation. In the testing phase, the feature of each target image from the probe camera is first encoded by the sparse representation and then recovered in the feature space spanned by the images from the gallery camera. The features of the same person from different cameras are similar following the above transformation. Experimental results on publicly available datasets demonstrate the superiority of our method.

****

What are You Talking About? Text-to-Image Coreference

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, Sanja Fidler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3558-3565

In this paper we exploit natural sentential descriptions of RGB-D scenes in order to improve 3D semantic parsing. Importantly, in doing so, we reason about which particular object each noun/pronoun is referring to in the image. This allows us to utilize visual information in order to disambiguate the so-called core

ference resolution problem that arises in text. Towards this goal, we propose a structure prediction model that exploits potentials computed from text and RGB-D imagery to reason about the class of the 3D objects, the scene type, as well as to align the nouns/pronouns with the referred visual objects. We demonstrate the effectiveness of our approach on the challenging NYU-RGBD v2 dataset, which we enrich with natural lingual descriptions. We show that our approach significantly improves 3D detection and scene classification accuracy, and is able to reliably estimate the text-to-image alignment. Furthermore, by using textual and visual information, we are also able to successfully deal with coreference in text, improving upon the state-of-the-art Stanford coreference system.
****
Predicting Failures of Vision Systems
Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3566-3573
Computer vision systems today fail frequently. They also fail abruptly without warning or explanation. Alleviating the former has been the primary focus of the community. In this work, we hope to draw the community's attention to the latter, which is arguably equally problematic for real applications. We promote two metrics to evaluate failure prediction. We show that a surprisingly straightforward and general approach, that we call ALERT, can predict the likely accuracy (or failure) of a variety of computer vision systems â■■ semantic segmentation, vanishing point and camera parameter estimation, and image memorability prediction â■■ on individual input images. We also explore attribute prediction, where classifiers are typically meant to generalize to new unseen categories. We show that ALERT can be useful in predicting failures of this transfer. Finally, we leverage ALERT to improve the performance of a downstream application of attribute prediction: zero-shot learning. We show that ALERT can outperform several strong baselines for zero-shot learning on four datasets.
****
Three Guidelines of Online Learning for Large-Scale Visual Recognition
Yoshitaka Ushiku, Masatoshi Hidaka, Tatsuya Harada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3574-3581
In this paper, we would like to evaluate online learning algorithms for large-scale visual recognition using state-of-the-art features which are preselected and held fixed. Today, combinations of high-dimensional features and linear classifiers are widely used for large-scale visual recognition. Numerous so-called mid-level features have been developed and mutually compared on an experimental basis. Although various learning methods for linear classification have also been proposed in the machine learning and natural language processing literature, they have rarely been evaluated for visual recognition. Therefore, we give guidelines via investigations of state-of-the-art online learning methods of linear classifiers. Many methods have been evaluated using toy data and natural language processing problems such as document classification. Consequently, we gave those methods a unified interpretation from the viewpoint of visual recognition. Results of controlled comparisons indicate three guidelines that might change the pipeline for visual recognition.
****
Using k-Poselets for Detecting People and Localizing Their Keypoints
Georgia Gkioxari, Bharath Hariharan, Ross Girshick, Jitendra Malik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3582-3589
A k-poselet is a deformable part model (DPM) with k parts, where each of the parts is a poselet, aligned to a specific configuration of keypoints based on ground-truth annotations. A separate template is used to learn the appearance of each part. The parts are allowed to move with respect to each other with a deformation cost that is learned at training time. This model is richer than both the traditional version of poselets and DPMs. It enables a unified approach to person detection and keypoint prediction which, barring contemporaneous approaches based on CNN features, achieves state-of-the-art keypoint prediction while maintainin

g competitive detection performance.

****

Randomized Max-Margin Compositions for Visual Recognition

Angela Eigenstetter, Masato Takami, Bjorn Ommer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3590-3597

A main theme in object detection are currently discriminative part-based models. The powerful model that combines all parts is then typically only feasible for few constituents, which are in turn iteratively trained to make them as strong as possible. We follow the opposite strategy by randomly sampling a large number of instance specific part classifiers. Due to their number, we cannot directly train a powerful classifier to combine all parts. Therefore, we randomly group them into fewer, overlapping compositions that are trained using a maximum-margin approach. In contrast to the common rationale of compositional approaches, we do not aim for semantically meaningful ensembles. Rather we seek randomized compositions that are discriminative and generalize over all instances of a category. Our approach not only localizes objects in cluttered scenes, but also explains them by parsing with compositions and their constituent parts. We conducted experiments on PASCAL VOC07, on the VOC10 evaluation server, and on the MITIndoor scene dataset. To the best of our knowledge, our randomized max-margin compositions (RM2C) are the currently best performing single class object detector using only HOG features. Moreover, the individual contributions of compositions and their parts are evaluated in separate experiments that demonstrate their potential.

****

Large-Scale Visual Font Recognition

Guang Chen, Jianchao Yang, Hailin Jin, Jonathan Brandt, Eli Shechtman, Aseem Agarwala, Tony X. Han; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3598-3605

This paper addresses the large-scale visual font recognition (VFR) problem, which aims at automatic identification of the typeface, weight, and slope of the text in an image or photo without any knowledge of content. Although visual font recognition has many practical applications, it has largely been neglected by the vision community. To address the VFR problem, we construct a large-scale dataset containing 2,420 font classes, which easily exceeds the scale of most image categorization datasets in computer vision. As font recognition is inherently dynamic and open-ended, i.e., new classes and data for existing categories are constantly added to the database over time, we propose a scalable solution based on the nearest class mean classifier (NCM). The core algorithm is built on local feature embedding, local feature metric learning and max-margin template selection, which is naturally amenable to NCM and thus to such open-ended classification problems. The new algorithm can generalize to new classes and new data at little added cost. Extensive experiments demonstrate that our approach is very effective on our synthetic test images, and achieves promising results on real world test images.

****

Describing Textures in the Wild

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, Andrea Vedaldi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3606-3613

Patterns and textures are key characteristics of many natural objects: a shirt can be striped, the wings of a butterfly can be veined, and the skin of an animal can be scaly. Aiming at supporting this dimension in image understanding, we address the problem of describing textures with semantic attributes. We identify a vocabulary of forty-seven texture terms and use them to describe a large dataset of patterns collected "in the wild". The resulting Describable Textures Dataset (DTD) is a basis to seek the best representation for recognizing describable texture attributes in images. We port from object recognition to texture recognition the Improved Fisher Vector (IFV) and Deep Convolutional-network Activation Features (DeCAF), and show that surprisingly, they both outperform specialized texture descriptors not only on our problem, but also in established material reco gnition datasets. We also show that our describable attributes are excellent texture descriptors, transferring between datasets and tasks; in particular, combined with IFV and DeCAF, they significantly outperform the state-of-the-art by more than 10% on both FMD and KTH-TIPS-2b benchmarks. We also demonstrate that they produce intuitive descriptions of materials and Internet images.

****

Relative Parts: Distinctive Parts for Learning Relative Attributes

Ramachandruni N. Sandeep, Yashaswi Verma, C. V. Jawahar; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3614-3621

The notion of relative attributes as introduced by Parikh and Grauman (ICCV, 2011) provides an appealing way of comparing two images based on their visual properties (or attributes) such as "smiling" for face images, "naturalness" for outdoor images, etc. For learning such attributes, a Ranking SVM based formulation was proposed that uses globally represented pairs of annotated images. In this paper, we extend this idea towards learning relative attributes using local parts that are shared across categories. First, instead of using a global representation, we introduce a part-based representation combining a pair of images that specifically compares corresponding parts. Then, with each part we associate a locally adaptive "significance-coefficient" that represents its discriminative ability with respect to a particular attribute. For each attribute, the significance-coefficients are learned simultaneously with a max-margin ranking model in an iterative manner. Compared to the baseline method, the new method is shown to achieve significant improvement in relative attribute prediction accuracy. Additionally, it is also shown to improve relative feedback based interactive image search.

****

Understanding Objects in Detail with Fine-Grained Attributes

Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B. Blaschko, David Weiss, Ben Taskar, Karen Simonyan, Naomi Saphra, Sammy Mohamed; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3622-3629

We study the problem of understanding objects in detail, intended as recognizing a wide array of fine-grained object attributes. To this end, we introduce a dataset of 7,413 airplanes annotated in detail with parts and their attributes, leveraging images donated by airplane spotters and crowdsourcing both the design and collection of the detailed annotations. We provide a number of insights that should help researchers interested in designing fine-grained datasets for other basic level categories. We show that the collected data can be used to study the relation between part detection and attribute prediction by diagnosing the performance of classifiers that pool information from different parts of an object. We note that the prediction of certain attributes can benefit substantially from accurate part detection. We also show that, differently from previous results in object detection, employing a large number of part templates can improve detection accuracy at the expenses of detection speed. We finally propose a coarse-to-fine approach to speed up detection through a hierarchical cascade algorithm.

****

Predicting User Annoyance Using Visual Attributes

Gordon Christie, Amar Parkash, Ujwal Krothapalli, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3630-3637

Computer Vision algorithms make mistakes. In human-centric applications, some mistakes are more annoying to users than others. In order to design algorithms that minimize the annoyance to users, we need access to an annoyance or cost matrix that holds the annoyance of each type of mistake. Such matrices are not readily available, especially for a wide gamut of human-centric applications where annoyance is tied closely to human perception. To avoid having to conduct extensive user studies to gather the annoyance matrix for all possible mistakes, we propose predicting the annoyance of previously unseen mistakes by learning from exampl e mistakes and their corresponding annoyance. We promote the use of attribute-based representations to transfer this knowledge of annoyance. Our experimental results with faces and scenes demonstrate that our approach can predict annoyance more accurately than baselines. We show that as a result, our approach makes less annoying mistakes in a real-world image retrieval application.

****

Linear Ranking Analysis

Weihong Deng, Jiani Hu, Jun Guo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3638-3645

We extend the classical linear discriminant analysis (LDA) technique to linear ranking analysis (LRA), by considering the ranking order of classes centroids on the projected subspace. Under the constrain on the ranking order of the classes, two criteria are proposed: 1) minimization of the classification error with the assumption that each class is homogenous Guassian distributed; 2) maximization of the sum (average) of the K minimum distances of all neighboring-class (centroid) pairs. Both criteria can be efficiently solved by the convex optimization for one-dimensional subspace. Greedy algorithm is applied to extend the results to the multi-dimensional subspace. Experimental results show that 1) LRA with both criteria achieve state-of-the-art performance on the tasks of ranking learning and zero-shot learning; and 2) the maximum margin criterion provides a discriminative subspace selection method, which can significantly remedy the class separation problem in comparing with several representative extensions of LDA.

****

Transformation Pursuit for Image Classification

Mattis Paulin, Jerome Revaud, Zaid Harchaoui, Florent Perronnin, Cordelia Schmid; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3646-3653

A simple approach to learning invariances in image clas- sification consists in augmenting the training set with transformed versions of the original images. However, given a large set of possible transformations, selecting a com- pact subset is challenging. Indeed, all transformations are not equally informative and adding uninformative transfor- mations increases training time with no gain in accuracy. We propose a principled algorithm â■■ Image Transformation Pursuit (ITP) â■■ for the automatic selection of a compact set of transformations. ITP works in a greedy fashion, by se- lecting at each iteration the one that yields the highest accuracy gain. ITP also allows to efficiently explore complex transformations, that combine basic transformations. We report results on two public benchmarks: the CUB dataset of bird images and the ImageNet 2010 challenge. Using Fisher Vector representations, we achieve an improvement from 28.2% to 45.2% in top-1 accuracy on CUB, and an im- provement from 70.1% to 74.9% in top-5 accuracy on Im- ageNet. We also show significant improvements for deep convnet features: from 47.3% to 55.4% on CUB and from 77.9% to 81.4% on ImageNet.

****

Incremental Learning of NCM Forests for Large-Scale Image Classification

Marko Ristin, Matthieu Guillaumin, Juergen Gall, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3654-3661

In recent years, large image data sets such as "ImageNet", "TinyImages" or ever-growing social networks like "Flickr" have emerged, posing new challenges to image classification that were not apparent in smaller image sets. In particular, the efficient handling of dynamically growing data sets, where not only the amount of training images, but also the number of classes increases over time, is a relatively unexplored problem. To remedy this, we introduce Nearest Class Mean Forests (NCMF), a variant of Random Forests where the decision nodes are based on nearest class mean (NCM) classification. NCMFs not only outperform conventional random forests, but are also well suited for integrating new classes. To this end, we propose and compare several approaches to incorporate data from new classes, so as to seamlessly extend the previously trained forest instead of re-training them from scratch. In our experiments, we show that NCMFs trained on small data sets with 10 classes can be extended to large data sets with 1000 classes wit hout significant loss of accuracy compared to training from scratch on the full data.

****

Object Classification with Adaptable Regions

Hakan Bilen, Marco Pedersoli, Vinay P. Namboodiri, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3662-3669

In classification of objects substantial work has gone into improving the low level representation of an image by considering various aspects such as different features, a number of feature pooling and coding techniques and considering different kernels. Unlike these works, in this paper, we propose to enhance the semantic representation of an image. We aim to learn the most important visual components of an image and how they interact in order to classify the objects correctly. To achieve our objective, we propose a new latent SVM model for category level object classification. Starting from image-level annotations, we jointly learn the object class and its context in terms of spatial location (where) and appearance (what). Furthermore, to regularize the complexity of the model we learn the spatial and co-occurrence relations between adjacent regions, such that unlikely configurations are penalized. Experimental results demonstrate that the proposed method can consistently enhance results on the challenging Pascal VOC dataset in terms of classification and weakly supervised detection. We also show how semantic representation can be exploited for finding similar content.

****

Discriminative Ferns Ensemble for Hand Pose Recognition

Eyal Krupka, Alon Vinnikov, Ben Klein, Aharon Bar Hillel, Daniel Freedman, Simon Stachniak; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3670-3677

We present the Discriminative Ferns Ensemble (DFE) classifier for efficient visual object recognition. The classifier architecture is designed to optimize both classification speed and accuracy when a large training set is available. Speed is obtained using simple binary features and direct indexing into a set of tables, and accuracy by using a large capacity model and careful discriminative optimization. The proposed framework is applied to the problem of hand pose recognition in depth and infra-red images, using a very large training set. Both the accuracy and the classification time obtained are considerably superior to relevant competing methods, allowing one to reach accuracy targets with run times orders of magnitude faster than the competition. We show empirically that using DFE, we can significantly reduce classification time by increasing training sample size for a fixed target accuracy. Finally a DFE result is shown for the MNIST dataset, showing the method's merit extends beyond depth images.

****

Are Cars Just 3D Boxes? - Jointly Estimating the 3D Shape of Multiple Objects

Muhammad Zeeshan Zia, Michael Stark, Konrad Schindler; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3678-3685

Current systems for scene understanding typically represent objects as 2D or 3D bounding boxes. While these representations have proven robust in a variety of applications, they provide only coarse approximations to the true 2D and 3D extent of objects. As a result, object-object interactions, such as occlusions or ground-plane contact, can be represented only superficially. In this paper, we approach the problem of scene understanding from the perspective of 3D shape modeling, and design a 3D scene representation that reasons jointly about the 3D shape of multiple objects. This representation allows to express 3D geometry and occlusion on the fine detail level of individual vertices of 3D wireframe models, and makes it possible to treat dependencies between objects, such as occlusion reasoning, in a deterministic way. In our experiments, we demonstrate the benefit of jointly estimating the 3D shape of multiple objects in a scene over working with coarse boxes, on the recently proposed KITTI dataset of realistic street scenes.

****

2D Human Pose Estimation: New Benchmark and State of the Art Analysis

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3686-3693

Human pose estimation has made significant progress during the last years. However current datasets are limited in their coverage of the overall pose estimation challenges. Still these serve as the common sources to evaluate, train and compare different models on. In this paper we introduce a novel benchmark "MPII Human Pose" that makes a significant advance in terms of diversity and difficulty, a contribution that we feel is required for future developments in human body models. This comprehensive dataset was collected using an established taxonomy of over 800 human activities. The collected images cover a wider variety of human activities than previous datasets including various recreational, occupational and householding activities, and capture people from a wider range of viewpoints. We provide a rich set of labels including positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels. For each image we provide adjacent video frames to facilitate the use of motion information. Given these rich annotations we perform a detailed analysis of leading human pose estimation approaches and gaining insights for the success and failures of these methods.

****

Using a Deformation Field Model for Localizing Faces and Facial Points under Weak Supervision

Marco Pedersoli, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3694-3701

Face detection and facial points localization are interconnected tasks. Recently it has been shown that solving these two tasks jointly with a mixture of trees of parts (MTP) leads to state-of-the-art results. However, MTP, as most other methods for facial point localization proposed so far, requires a complete annotation of the training data at facial point level. This is used to predefine the structure of the trees and to place the parts correctly. In this work we extend the mixtures from trees to more general loopy graphs. In this way we can learn in a weakly supervised manner (using only the face location and orientation) a powerful deformable detector that implicitly aligns its parts to the detected face in the image. By attaching some reference points to the correct parts of our detector we can then localize the facial points. In terms of detection our method clearly outperforms the state-of-the-art, even if competing with methods that use facial point annotations during training. Additionally, without any facial point annotation at the level of individual training images, our method can localize facial points with an accuracy similar to fully supervised approaches.

****

Active Annotation Translation

Steve Branson, Kristjan Eldjarn Hjorleifsson, Pietro Perona; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3702-3709

We introduce a general framework for quickly annotating an image dataset when previous annotations exist. The new annotations (e.g. part locations) may be quite different from the old annotations (e.g. segmentations). Human annotators may be thought of as helping translate the old annotations into the new ones. As annotators label images, our algorithm incrementally learns a translator from source to target labels as well as a computer-vision-based structured predictor. These two components are combined to form an improved prediction system which accelerates the annotators' work through a smart GUI.   We show how the method can be applied to translate between a wide variety of annotation types, including bounding boxes, segmentations, 2D and 3D part-based systems, and class and attribute labels. The proposed system will be a useful tool toward exploring new types of representations beyond simple bounding boxes, object segmentations, and class labels, and toward finding new ways to exploit existing large datasets with traditional types of annotations like SUN, Image Net, and Pascal VOC. Experiments on the CUB-200-2011 and H3D datasets demonstrate 1) our method accelerates collection of part annotations by a factor of 3-20 compared to manual labeling, 2) our system can be used effectively in a scheme where definitions of part, attribute, or action vocabularies are evolved interactively without relabeling the entire dataset, and 3) toward collecting pose annotations, segmentations are more useful than bounding boxes, and part-level annotations are more effective than segmentations.

****

Looking Beyond the Visible Scene

Aditya Khosla, Byoungkwon An An, Joseph J. Lim, Antonio Torralba; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3710-3717

A common thread that ties together many prior works in scene understanding is their focus on the aspects directly present in a scene such as its categorical classification or the set of objects. In this work, we propose to look beyond the visible elements of a scene; we demonstrate that a scene is not just a collection of objects and their configuration or the labels assigned to its pixels - it is so much more. From a simple observation of a scene, we can tell a lot about the environment surrounding the scene such as the potential establishments near it, the potential crime rate in the area, or even the economic climate. Here, we explore several of these aspects from both the human perception and computer vision perspective. Specifically, we show that it is possible to predict the distance of surrounding establishments such as McDonaldâ■■s or hospitals even by using scenes located far from them. We go a step further to show that both humans and computers perform well at navigating the environment based only on visual cues from scenes. Lastly, we show that it is possible to predict the crime rates in an area simply by looking at a scene without any real-time criminal activity. Simply put, here, we illustrate that it is possible to look beyond the visible scene.

****

Two-Class Weather Classification

Cewu Lu, Di Lin, Jiaya Jia, Chi-Keung Tang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3718-3725

Given a single outdoor image, this paper proposes a collaborative learning approach for labeling it as either sunny or cloudy. Never adequately addressed, this twoclass classification problem is by no means trivial given the great variety of outdoor images. Our weather feature combines special cues after properly encoding them into feature vectors. They then work collaboratively in synergy under a unified optimization framework that is aware of the presence (or absence) of a given weather cue during learning and classification. Extensive experiments and comparisons are performed to verify our method. We build a new weather image dataset consisting of 10K sunny and cloudy images, which is available online together with the executable.

****

Learning Important Spatial Pooling Regions for Scene Classification

Di Lin, Cewu Lu, Renjie Liao, Jiaya Jia; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3726-3733

We address the false response influence problem when learning and applying discriminative parts to construct the mid-level representation in scene classification. It is often caused by the complexity of latent image structure when convolving part filters with input images. This problem makes mid-level representation, even after pooling, not distinct enough to classify input data correctly to categories. Our solution is to learn important spatial pooling regions along with their appearance. The experiments show that this new framework suppresses false response and produces improved results on several datasets, including MIT-Indoor, 15-Scene, and UIUC 8-Sport. When combined with global image features, our method achieves state-of-the-art performance on these datasets.

****

Orientational Pyramid Matching for Recognizing Indoor Scenes

Lingxi Xie, Jingdong Wang, Baining Guo, Bo Zhang, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3734-3741

Scene recognition is a basic task towards image understanding. Spatial Pyramid M atching (SPM) has been shown to be an efficient solution for spatial context modeling. In this paper, we introduce an alternative approach, Orientational Pyramid Matching (OPM), for orientational context modeling. Our approach is motivated by the observation that the 3D orientations of objects are a crucial factor to discriminate indoor scenes. The novelty lies in that OPM uses the 3D orientations to form the pyramid and produce the pooling regions, which is unlike SPM that uses the spatial positions to form the pyramid. Experimental results on challenging scene classification tasks show that OPM achieves the performance comparable with SPM and that OPM and SPM make complementary contributions so that their combination gives the state-of-the-art performance.

****

Multilabel Ranking with Inconsistent Rankers

Xin Geng, Longrun Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3742-3747

While most existing multilabel ranking methods assume the availability of a single objective label ranking for each instance in the training set, this paper deals with a more common case where subjective inconsistent rankings from multiple rankers are associated with each instance. The key idea is to learn a latent preference distribution for each instance. The proposed method mainly includes two steps. The first step is to generate a common preference distribution that is most compatible to all the personal rankings. The second step is to learn a mapping from the instances to the preference distributions. The proposed preference distribution learning (PDL) method is applied to the problem of multilabel ranking for natural scene images. Experimental results show that PDL can effectively incorporate the information given by the inconsistent rankers, and perform remarkably better than the compared state-of-the-art multilabel ranking algorithms.

****

Scene Parsing with Object Instances and Occlusion Ordering

Joseph Tighe, Marc Niethammer, Svetlana Lazebnik; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3748-3755

This work proposes a method to interpret a scene by assigning a semantic label at every pixel and inferring the spatial extent of individual object instances together with their occlusion relationships. Starting with an initial pixel labeling and a set of candidate object masks for a given test image, we select a subset of objects that explain the image well and have valid overlap relationships and occlusion ordering. This is done by minimizing an integer quadratic program either using a greedy method or a standard solver. Then we alternate between using the object predictions to refine the pixel labels and vice versa. The proposed system obtains promising results on two challenging subsets of the LabelMe and SUN datasets, the largest of which contains 45,676 images and 232 classes.

****

A Riemannian Framework for Matching Point Clouds Represented by the Schrodinger Distance Transform

Yan Deng, Anand Rangarajan, Stephan Eisenschenk, Baba C. Vemuri; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3756-3761

In this paper, we cast the problem of point cloud matching as a shape matching problem by transforming each of the given point clouds into a shape representation called the Schrodinger distance transform (SDT) representation. This is achieved by solving a static Schrodinger equation instead of the corresponding static Hamilton-Jacobi equation in this setting. The SDT representation is an analytic expression and following the theoretical physics literature, can be normalized to have unit L_2 norm---making it a square-root density, which is identified with a point on a unit Hilbert sphere, whose intrinsic geometry is fully known. The Fisher-Rao metric, a natural metric for the space of densities leads to analytic expressions for the geodesic distance between points on this sphere. In this paper, we use the well known Riemannian framework never before used for point cloud matching, and present a novel matching algorithm. We pose point set matching under rigid and non-rigid transformations in this framework and solve for the transformations using standard nonlinear optimization techniques. Finally, to eva luate the performance of our algorithm---dubbed SDTM---we present several synthetic and real data examples along with extensive comparisons to state-of-the-art techniques. The experiments show that our algorithm outperforms state-of the-art point set registration algorithms on many quantitative metrics.

****

Seeing 3D Chairs: Exemplar Part-based 2D-3D Alignment using a Large Dataset of CAD Models

Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, Josef Sivic; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3762-3769

This paper poses object category detection in images as a type of 2D-to-3D alignment problem, utilizing the large quantities of 3D CAD models that have been made publicly available online. Using the "chair" class as a running example, we propose an exemplar-based 3D category representation, which can explicitly model chairs of different styles as well as the large variation in viewpoint. We develop an approach to establish part-based correspondences between 3D CAD models and real photographs. This is achieved by (i) representing each 3D model using a set of view-dependent mid-level visual elements learned from synthesized views in a discriminative fashion, (ii) carefully calibrating the individual element detectors on a common dataset of negative images, and (iii) matching visual elements to the test image allowing for small mutual deformations but preserving the viewpoint and style constraints. We demonstrate the ability of our system to align 3D models with 2D objects in the challenging PASCAL VOC images, which depict a wide variety of chairs in complex scenes.

****

A Mixture of Manhattan Frames: Beyond the Manhattan World

Julian Straub, Guy Rosman, Oren Freifeld, John J. Leonard, John W. Fisher III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3770-3777

Objects and structures within man-made environments typically exhibit a high degree of organization in the form of orthogonal and parallel planes. Traditional approaches to scene representation exploit this phenomenon via the somewhat restrictive assumption that every plane is perpendicular to one of the axes of a single coordinate system. Known as the Manhattan-World model, this assumption is widely used in computer vision and robotics. The complexity of many real-world scenes, however, necessitates a more flexible model. We propose a novel probabilistic model that describes the world as a mixture of Manhattan frames: each frame defines a different orthogonal coordinate system. This results in a more expressive model that still exploits the orthogonality constraints. We propose an adaptive Markov-Chain Monte-Carlo sampling algorithm with Metropolis-Hastings split/merge moves that utilizes the geometry of the unit sphere. We demonstrate the versatility of our Mixture-of-Manhattan-Frames model by describing complex scenes using depth images of indoor scenes as well as aerial-LiDAR measurements of an urban center. Additionally, we show that the model lends itself to focal-length calibration of depth cameras and to plane segmentation.

****

Local Regularity-driven City-scale Facade Detection from Aerial Images

Jingchen Liu, Yanxi Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3778-3785

We propose a novel regularity-driven framework for facade detection from aerial images of urban scenes. Gini-index is used in our work to form an edge-based regularity metric relating regularity and distribution sparsity. Facade regions are chosen so that these local regularities are maximized. We apply a greedy adaptive region expansion procedure for facade region detection and growing, followed by integer quadratic programming for removing overlapping facades to optimize facade coverage. Our algorithm can handle images that have wide viewing angles and contain more than 200 facades per image. The experimental results on images from three different cities (NYC, Rome, San-Francisco) demonstrate superior performance on facade detection in both accuracy and speed over state of the art methods. We also show an application of our facade detection for effective cross-view

facade matching.
****

Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture
Danhang Tang, Hyung Jin Chang, Alykhan Tejani, Tae-Kyun Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3786-3793

In this paper we present the Latent Regression Forest (LRF), a novel framework for real-time, 3D hand pose estimation from a single depth image. In contrast to prior forest-based methods, which take dense pixels as input, classify them independently and then estimate joint positions afterwards; our method can be considered as a structured coarse-to-fine search, starting from the centre of mass of a point cloud until locating all the skeletal joints. The searching process is guided by a learnt Latent Tree Model which reflects the hierarchical topology of the hand. Our main contributions can be summarised as follows: (i) Learning the topology of the hand in an unsupervised, data-driven manner. (ii) A new forest-based, discriminative framework for structured search in images, as well as an error regression step to avoid error accumulation. (iii) A new multi-view hand pose dataset containing 180K annotated images from 10 different subjects. Our experiments show that the LRF out-performs state-of-the-art methods in both accuracy and efficiency.
****

FAUST: Dataset and Evaluation for 3D Mesh Registration
Federica Bogo, Javier Romero, Matthew Loper, Michael J. Black; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3794-3801

New scanning technologies are increasing the importance of 3D mesh data and the need for algorithms that can reliably align it. Surface registration is important for building full 3D models from partial scans, creating statistical shape models, shape retrieval, and tracking. The problem is particularly challenging for non-rigid and articulated objects like human bodies. While the challenges of real-world data registration are not present in existing synthetic datasets, establishing ground-truth correspondences for real 3D scans is difficult. We address this with a novel mesh registration technique that combines 3D shape and appearance information to produce high-quality alignments. We define a new dataset called FAUST that contains 300 scans of 10 people in a wide range of poses together with an evaluation methodology. To achieve accurate registration, we paint the subjects with high-frequency textures and use an extensive validation process to ensure accurate ground truth. We find that current shape registration methods have trouble with this real-world data. The dataset and evaluation website are available for research purposes at http://faust.is.tue.mpg.de.
****

Optimizing Over Radial Kernels on Compact Manifolds
Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, Mehrtash Harandi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3802-3809

We tackle the problem of optimizing over all possible positive definite radial kernels on Riemannian manifolds for classification. Kernel methods on Riemannian manifolds have recently become increasingly popular in computer vision. However, the number of known positive definite kernels on manifolds remain very limited. Furthermore, most kernels typically depend on at least one parameter that needs to be tuned for the problem at hand. A poor choice of kernel, or of parameter value, may yield significant performance drop-off. Here, we show that positive definite radial kernels on the unit n-sphere, the Grassmann manifold and Kendall's shape manifold can be expressed in a simple form whose parameters can be automatically optimized within a support vector machine framework. We demonstrate the benefits of our kernel learning algorithm on object, face, action and shape recognition.
****

Grassmann Averages for Scalable Robust PCA
Soren Hauberg, Aasa Feragen, Michael J. Black; Proceedings of the IEEE Conferenc

e on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3810-3817

As the collection of large datasets becomes increasingly automated, the occurrence of outliers will increase -- "big data" implies "big outliers''. While principal component analysis (PCA) is often used to reduce the size of data, and scalable solutions exist, it is well-known that outliers can arbitrarily corrupt the results. Unfortunately, state-of-the-art approaches for robust PCA do not scale beyond small-to-medium sized datasets. To address this, we introduce the Grassmann Average (GA), which expresses dimensionality reduction as an average of the subspaces spanned by the data. Because averages can be efficiently computed, we immediately gain scalability. GA is inherently more robust than PCA, but we show that they coincide for Gaussian data. We exploit that averages can be made robust to formulate the Robust Grassmann Average (RGA) as a form of robust PCA. Robustness can be with respect to vectors (subspaces) or elements of vectors; we focus on the latter and use a trimmed average. The resulting Trimmed Grassmann Average (TGA) is particularly appropriate for computer vision because it is robust to pixel outliers. The algorithm has low computational complexity and minimal memory requirements, making it scalable to "big noisy data." We demonstrate TGA for background modeling, video restoration, and shadow removal. We show scalability by performing robust PCA on the entire Star Wars IV movie.

****

Robust Subspace Segmentation with Block-diagonal Prior
Jiashi Feng, Zhouchen Lin, Huan Xu, Shuicheng Yan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3818-3825

The subspace segmentation problem is addressed in this paper by effectively constructing an exactly block-diagonal sample affinity matrix. The block-diagonal structure is heavily desired for accurate sample clustering but is rather difficult to obtain. Most current state-of-the-art subspace segmentation methods (such as SSC and LRR) resort to alternative structural priors (such as sparseness and low-rankness) to construct the affinity matrix. In this work, we directly pursue the block-diagonal structure by proposing a graph Laplacian constraint based formulation, and then develop an efficient stochastic subgradient algorithm for optimization. Moreover, two new subspace segmentation methods, the block-diagonal SSC and LRR, are devised in this work. To the best of our knowledge, this is the first research attempt to explicitly pursue such a block-diagonal structure. Extensive experiments on face clustering, motion segmentation and graph construction for semi-supervised learning clearly demonstrate the superiority of our novelly proposed subspace segmentation methods.

****

Unsupervised One-Class Learning for Automatic Outlier Removal
Wei Liu, Gang Hua, John R. Smith; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3826-3833

Outliers are pervasive in many computer vision and pattern recognition problems. Automatically eliminating outliers scattering among practical data collections becomes increasingly important, especially for Internet inspired vision applications. In this paper, we propose a novel one-class learning approach which is robust to contamination of input training data and able to discover the outliers that corrupt one class of data source. Our approach works under a fully unsupervised manner, differing from traditional one-class learning supervised by known positive labels. By design, our approach optimizes a kernel-based max-margin objective which jointly learns a large margin one-class classifier and a soft label assignment for inliers and outliers. An alternating optimization algorithm is then designed to iteratively refine the classifier and the labeling, achieving a provably convergent solution in only a few iterations. Extensive experiments conducted on four image datasets in the presence of artificial and real-world outliers demonstrate that the proposed approach is considerably superior to the state-of-the-arts in obliterating outliers from contaminated one class of images, exhibiting strong robustness at a high outlier proportion up to 60%.

****

Smooth Representation Clustering
Han Hu, Zhouchen Lin, Jianjiang Feng, Jie Zhou; Proceedings of the IEEE Conferen

ce on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3834-3841

Subspace clustering is a powerful technology for clustering data according to the underlying subspaces. Representation based methods are the most popular subspace clustering approach in recent years. In this paper, we analyze the grouping effect of representation based methods in depth. In particular, we introduce the enforced grouping effect conditions, which greatly facilitate the analysis of grouping effect. We further find that grouping effect is important for subspace clustering, which should be explicitly enforced in the data self-representation model, rather than implicitly implied by the model as in some prior work. Based on our analysis, we propose the SMooth Representation (SMR) model. We also propose a new affinity measure based on the grouping effect, which proves to be much more effective than the commonly used one. As a result, our SMR significantly outperforms the state-of-the-art ones on benchmark datasets.

****

Novel Methods for Multilinear Data Completion and De-noising Based on Tensor-SVD

Zemin Zhang, Gregory Ely, Shuchin Aeron, Ning Hao, Misha Kilmer; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3842-3849

In this paper we propose novel methods for completion (from limited samples) and de-noising of multilinear (tensor) data and as an application consider 3-D and 4- D (color) video data completion and de-noising. We exploit the recently proposed tensor-Singular Value Decomposition (t-SVD)[11]. Based on t-SVD, the notion of multilinear rank and a related tensor nuclear norm was proposed in [11] to characterize informational and structural complexity of multilinear data. We first show that videos with linear camera motion can be represented more efficiently using t-SVD compared to the approaches based on vectorizing or flattening of the tensors. Since efficiency in representation implies efficiency in recovery, we outline a tensor nuclear norm penalized algorithm for video completion from missing entries. Application of the proposed algorithm for video recovery from missing entries is shown to yield a superior performance over existing methods. We also consider the problem of tensor robust Principal Component Analysis (PCA) for de-noising 3-D video data from sparse random corruptions. We show superior performance of our method compared to the matrix robust PCA adapted to this setting as proposed in [4].

****

Second-Order Shape Optimization for Geometric Inverse Problems in Vision

Jonathan Balzer, Stefano Soatto; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3850-3857

We develop a method for optimization in shape spaces, i.e., sets of surfaces modulo re-parametrization. Unlike previously proposed gradient flows, we achieve superlinear convergence rates through an approximation of the shape Hessian, which is generally hard to compute and suffers from a series of degeneracies. Our analysis highlights the role of mean curvature motion in comparison with first-order schemes: instead of surface area, our approach penalizes deformation, either by its Dirichlet energy or total variation, and hence does not suffer from shrinkage. The latter regularizer sparks the development of an alternating direction method of multipliers on triangular meshes. Therein, a conjugate-gradient solver enables us to bypass formation of the Gaussian normal equations appearing in the course of the overall optimization. We combine all of these ideas in a versatile geometric variation-regularized Levenberg-Marquardt-type method applicable to a variety of shape functionals, depending on intrinsic properties of the surface such as normal field and curvature as well as its embedding into space. Promising experimental results are reported.

****

l0 Norm Based Dictionary Learning by Proximal Methods with Global Convergence

Chenglong Bao, Hui Ji, Yuhui Quan, Zuowei Shen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3858-3865

Sparse coding and dictionary learning have seen their applications in many vision tasks, which usually is formulated as a non-convex optimization problem. Many iterative methods have been proposed to tackle such an optimization problem. How ever, it remains an open problem to have a method that is not only practically fast but also is globally convergent. In this paper, we proposed a fast proximal method for solving l0 norm based dictionary learning problems, and we proved that the whole sequence generated by the proposed method converges to a stationary point with sub-linear convergence rate. The benefit of having a fast and convergent dictionary learning method is demonstrated in the applications of image recovery and face recognition.

****

Adaptive Partial Differential Equation Learning for Visual Saliency Detection

Risheng Liu, Junjie Cao, Zhouchen Lin, Shiguang Shan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3866-3873

Partial Differential Equations (PDEs) have been successful in solving many low-level vision tasks. However, it is a challenging task to directly utilize PDEs for visual saliency detection due to the difficulty in incorporating human perception and high-level priors to a PDE system. Instead of designing PDEs with fixed formulation and boundary condition, this paper proposes a novel framework for adaptively learning a PDE system from an image for visual saliency detection. We assume that the saliency of image elements can be carried out from the relevances to the saliency seeds (i.e., the most representative salient elements). In this view, a general Linear Elliptic System with Dirichlet boundary (LESD) is introduced to model the diffusion from seeds to other relevant points. For a given image, we first learn a guidance map to fuse human prior knowledge to the diffusion system. Then by optimizing a discrete submodular function constrained with this LESD and a uniform matroid, the saliency seeds (i.e., boundary conditions) can be learnt for this image, thus achieving an optimal PDE system to model the evolution of visual saliency. Experimental results on various challenging image sets show the superiority of our proposed learning-based PDEs for visual saliency detection.

****

Robust Orthonormal Subspace Learning: Efficient Recovery of Corrupted Low-rank Matrices

Xianbiao Shu, Fatih Porikli, Narendra Ahuja; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3874-3881

Low-rank matrix recovery from a corrupted observation has many applications in computer vision. Conventional methods address this problem by iterating between nuclear norm minimization and sparsity minimization. However, iterative nuclear norm minimization is computationally prohibitive for large-scale data (e.g., video) analysis. In this paper, we propose a Robust Orthogonal Subspace Learning (ROSL) method to achieve efficient low-rank recovery. Our intuition is a novel rank measure on the low-rank matrix that imposes the group sparsity of its coefficients under orthonormal subspace. We present an efficient sparse coding algorithm to minimize this rank measure and recover the low-rank matrix at quadratic complexity of the matrix size. We give theoretical proof to validate that this rank measure is lower bounded by nuclear norm and it has the same global minimum as the latter. To further accelerate ROSL to linear complexity, we also describe a faster version (ROSL+) empowered by random sampling. Our extensive experiments demonstrate that both ROSL and ROSL+ provide superior efficiency against the state-of-the-art methods at the same level of recovery accuracy.

****

Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos

Gunhee Kim, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3882-3889

In this paper, we investigate an approach for reconstructing storyline graphs from large-scale collections of Internet images, and optionally other side information such as friendship graphs. The storyline graphs can be an effective summary that visualizes various branching narrative structure of events or activities recurring across the input photo sets of a topic class. In order to explore further the usefulness of the storyline graphs, we leverage them to perform the image sequential prediction tasks, from which photo recommendation applications can b enefit. We formulate the storyline reconstruction problem as an inference of sparse time-varying directed graphs, and develop an optimization algorithm that successfully addresses a number of key challenges of Web-scale problems, including global optimality, linear complexity, and easy parallelization. With experiments on more than 3.3 millions of images of 24 classes and user studies via Amazon Mechanical Turk, we show that the proposed algorithm improves other candidate methods for both storyline reconstruction and image prediction tasks.

****

Active Flattening of Curved Document Images via Two Structured Beams

Gaofeng Meng, Ying Wang, Shenquan Qu, Shiming Xiang, Chunhong Pan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3890-3897

Document images captured by a digital camera often suffer from serious geometric distortions. In this paper,we propose an active method to correct geometric distortions in a camera-captured document image. Unlike many passive rectification methods that rely on text-lines or features extracted from images, our method uses two structured beams illuminating upon the document page to recover two spatial curves. A developable surface is then interpolated to the curves by finding the correspondence between them. The developable surface is finally flattened onto a plane by solving a system of ordinary differential equations. Our method is a content independent approach and can restore a corrected document image of high accuracy with undistorted contents. Experimental results on a variety of real-captured document images demonstrate the effectiveness and efficiency of the proposed method.

****

Image-based Synthesis and Re-Synthesis of Viewpoints Guided by 3D Models

Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3898-3905

We propose a technique to use the structural information extracted from a set of 3D models of an object class to improve novel-view synthesis for images showing unknown instances of this class. These novel views can be used to "amplify" training image collections that typically contain only a low number of views or lack certain classes of views entirely (e.g. top views). We extract the correlation of position, normal, reflectance and appearance from computer-generated images of a few exemplars and use this information to infer new appearance for new instances. We show that our approach can improve performance of state-of-the-art detectors using real-world training data. Additional applications include guided versions of inpainting, 2D-to-3D conversion, super-resolution and non-local smoothing.

****

Bayesian View Synthesis and Image-Based Rendering Principles

Sergi Pujades, Frederic Devernay, Bastian Goldluecke; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3906-3913

In this paper, we address the problem of synthesizing novel views from a set of input images. State of the art methods, such as the Unstructured Lumigraph, have been using heuristics to combine information from the original views, often using an explicit or implicit approximation of the scene geometry. While the proposed heuristics have been largely explored and proven to work effectively, a Bayesian formulation was recently introduced, formalizing some of the previously proposed heuristics, pointing out which physical phenomena could lie behind each. However, some important heuristics were still not taken into account and lack proper formalization. We contribute a new physics-based generative model and the corresponding Maximum a Posteriori estimate, providing the desired unification between heuristics-based methods and a Bayesian formulation. The key point is to systematically consider the error induced by the uncertainty in the geometric proxy. We provide an extensive discussion, analyzing how the obtained equations explain the heuristics developed in previous methods. Furthermore, we show that our novel Bayesian model significantly improves the quality of novel views, in particular if the scene geometry estimate is inaccurate.

****

Fast MRF Optimization with Application to Depth Reconstruction

Qifeng Chen, Vladlen Koltun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3914-3921

We describe a simple and fast algorithm for optimizing Markov random fields over images. The algorithm performs block coordinate descent by optimally updating a horizontal or vertical line in each step. While the algorithm is not as accurate as state-of-the-art MRF solvers on traditional benchmark problems, it is trivially parallelizable and produces competitive results in a fraction of a second. As an application, we develop an approach to increasing the accuracy of consumer depth cameras. The presented algorithm enables high-resolution MRF optimization at multiple frames per second and substantially increases the accuracy of the produced range images.

****

Exploiting Shading Cues in Kinect IR Images for Geometry Refinement

Gyeongmin Choe, Jaesik Park, Yu-Wing Tai, In So Kweon; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3922-3929

In this paper, we propose a method to refine geometry of 3D meshes from the Kinect fusion by exploiting shading cues captured from the infrared (IR) camera of Kinect. A major benefit of using the Kinect IR camera instead of a RGB camera is that the IR images captured by Kinect are narrow band images which filtered out most undesired ambient light that makes our system robust to natural indoor illumination. We define a near light IR shading model which describes the captured intensity as a function of surface normals, albedo, lighting direction, and distance between a light source and surface points. To resolve ambiguity in our model between normals and distance, we utilize an initial 3D mesh from the Kinect fusion and multi-view information to reliably estimate surface details that were not reconstructed by the Kinect fusion. Our approach directly operates on a 3D mesh model for geometry refinement. The effectiveness of our approach is demonstrated through several challenging real-world examples.

****

Fast Rotation Search with Stereographic Projections for 3D Registration

Alvaro Parra Bustos, Tat-Jun Chin, David Suter; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3930-3937

Recently there has been a surge of interest to use branch-and-bound (bnb) optimisation for 3D point cloud registration. While bnb guarantees globally optimal solutions, it is usually too slow to be practical. A fundamental source of difficulty is the search for the rotation parameters in the 3D rigid transform. In this work, assuming that the translation parameters are known, we focus on constructing a fast rotation search algorithm. With respect to an inherently robust geometric matching criterion, we propose a novel bounding function for bnb that allows rapid evaluation. Underpinning our bounding function is the usage of stereographic projections to precompute and spatially index all possible point matches. This yields a robust and global algorithm that is significantly faster than previous methods. To conduct full 3D registration, the translation can be supplied by 3D feature matching, or by another optimisation framework that provides the translation. On various challenging point clouds, including those taken out of lab settings, our approach demonstrates superior efficiency.

****

Local Readjustment for High-Resolution 3D Reconstruction

Siyu Zhu, Tian Fang, Jianxiong Xiao, Long Quan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3938-3945

Global bundle adjustment usually converges to a non-zero residual and produces sub-optimal camera poses for local areas, which leads to loss of details for high- resolution reconstruction. Instead of trying harder to optimize everything globally, we argue that we should live with the non-zero residual and adapt the camera poses to local areas. To this end, we propose a segment-based approach to readjust the camera poses locally and improve the reconstruction for fine geometry details. The key idea is to partition the globally optimized structure from motion points into well-conditioned segments for re-optimization, reconstruct their geometry individually, and fuse everything back into a consistent global model. This significantly reduces severe propagated errors and estimation biases caused by the initial global adjustment. The results on several datasets demonstrate that this approach can significantly improve the reconstruction accuracy, while maintaining the consistency of the 3D structure between segments.

****

Turning Mobile Phones into 3D Scanners

Kalin Kolev, Petri Tanskanen, Pablo Speciale, Marc Pollefeys; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3946-3953

In this paper, we propose an efficient and accurate scheme for the integration of multiple stereo-based depth measurements. For each provided depth map a confidence-based weight is assigned to each depth estimate by evaluating local geometry orientation, underlying camera setting and photometric evidence. Subsequently, all hypotheses are fused together into a compact and consistent 3D model. Thereby, visibility conflicts are identified and resolved, and fitting measurements are averaged with regard to their confidence scores. The individual stages of the proposed approach are validated by comparing it to two alternative techniques which rely on a conceptually different fusion scheme and a different confidence inference, respectively. Pursuing live 3D reconstruction on mobile devices as a primary goal, we demonstrate that the developed method can easily be integrated into a system for monocular interactive 3D modeling by substantially improving its accuracy while adding a negligible overhead to its performance and retaining its interactive potential.

****

T-Linkage: A Continuous Relaxation of J-Linkage for Multi-Model Fitting

Luca Magri, Andrea Fusiello; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3954-3961

This paper presents an improvement of the J-linkage algorithm for fitting multiple instances of a model to noisy data corrupted by outliers. The binary preference analysis implemented by J-linkage is replaced by a continuous (soft, or fuzzy) generalization that proves to perform better than J-linkage on simulated data, and compares favorably with state of the art methods on public domain real datasets.

****

Motion-Depth: RGB-D Depth Map Enhancement with Motion and Depth in Complement

Tak-Wai Hui, King Ngi Ngan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3962-3969

Low-cost RGB-D imaging system such as Kinect is widely utilized for dense 3D reconstruction. However, RGB-D system generally suffers from two main problems. The spatial resolution of the depth image is low. The depth image often contains numerous holes where no depth measurements are available. This can be due to bad infra-red reflectance properties of some objects in the scene. Since the spatial resolution of the color image is generally higher than that of the depth image, this paper introduces a new method to enhance the depth images captured by a moving RGB-D system using the depth cues from the induced optical flow. We not only fill the holes in the raw depth images, but also recover fine details of the imaged scene. We address the problem of depth image enhancement by minimizing an energy functional. In order to reduce the computational complexity, we have treated the textured and homogeneous regions in the color images differently. Experimental results on several RGB-D sequences are provided to show the effectiveness of the proposed method.

****

Generalized Pupil-Centric Imaging and Analytical Calibration for a Non-frontal Camera

Avinash Kumar, Narendra Ahuja; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3970-3977

We consider the problem of calibrating a small field of view central perspective non-frontal camera whose lens and sensor planes may not be parallel to each other. This can be due to manufacturing defects or intentional tilting. Thus, as su

ch all cameras can be modeled as being non-frontal with varying degrees. There are two approaches to model non- frontal cameras. The first one based on rotation parameterization of sensor non-frontalness/tilt increases the number of calibration parameters, thus requiring heuristics to initialize a few calibration parameters for the final non-linear optimization step. Additionally, for this parameterization, while it has been shown that pupil-centric imaging model leads to more accurate rotation estimates than a thin-lens imaging model, it has only been developed for a single axis lens-sensor tilt. But, in real cameras we can have arbitrary tilt. The second approach based on decentering distortion modeling is approximate as it can only handle small tilts and cannot explicitly estimate the sensor tilt. In this paper, we focus on rotation based non-frontal camera calibration and address the aforementioned problems of over-parameterization and inadequacy of existing pupil-centric imaging model. We first derive a generalized pupil-centric imaging model for arbitrary axis lens-sensor tilt. We then derive an analytical solution, in this setting, for a subset of calibration parameters including sensor rotation angles as a function of center of radial distortion (CoD). A radial alignment based constraint is then proposed to computationally estimate CoD leveraging on the proposed analytical solution. Our analytical technique also estimates pupil-centric parameters of entrance pupil location and optical focal length, which have typically been done optically. Given these analytical and computational calibration parameter estimates, we initialize the non-linear calibration optimization for a set of synthetic and real data captured from a non-frontal camera and show reduced pixel re-projection and undistortion errors compared to state of the art techniques in rotation and decentering based approaches to non-frontal camera calibration.
****

Geometric Urban Geo-Localization
Mayank Bansal, Kostas Daniilidis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3978-3985
We propose a purely geometric correspondence-free approach to urban geo-localization using 3D point-ray features extracted from the Digital Elevation Map of an urban environment. We derive a novel formulation for estimating the camera pose locus using 3D-to-2D correspondence of a single point and a single direction alone. We show how this allows us to compute putative correspondences between building corners in the DEM and the query image by exhaustively combining pairs of point-ray features. Then, we employ the two-point method to estimate both the camera pose and compute correspondences between buildings in the DEM and the query image. Finally, we show that the computed camera poses can be efficiently ranked by a simple skyline projection step using building edges from the DEM. Our experimental evaluation illustrates the promise of a purely geometric approach to the urban geo-localization problem.
****

3D Reconstruction from Accidental Motion
Fisher Yu, David Gallup; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3986-3993
We have discovered that 3D reconstruction can be achieved from asingle still photographic capture due to accidental motions of thephotographer, even while attempting to hold the camera still. Although these motions result in little baseline and therefore high depth uncertainty, in theory, we can combine many such measurements over the duration of the capture process (a few seconds) to achieve usable depth estimates. Wepresent a novel 3D reconstruction system tailored for this problemthat produces depth maps from short video sequences from standard cameraswithout the need for multi-lens optics, active sensors, or intentionalmotions by the photographer. This result leads to the possibilitythat depth maps of sufficient quality for RGB-D photography applications likeperspective change, simulated aperture, and object segmentation, cancome "for free" for a significant fraction of still photographsunder reasonable conditions.
****

Real-time Model-based Articulated Object Pose Detection and Tracking with Variable Rigidity Constraints

Karl Pauwels, Leonardo Rubio, Eduardo Ros; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3994-4001

A novel model-based approach is introduced for real-time detection and tracking of the pose of general articulated objects. A variety of dense motion and depth cues are integrated into a novel articulated Iterative Closest Point approach. The proposed method can independently track the six-degrees-of-freedom pose of over a hundred of rigid parts in real-time while, at the same time, imposing articulation constraints on the relative motion of different parts. We propose a novel rigidization framework for optimally handling unobservable parts during tracking. This involves rigidly attaching the minimal amount of unseen parts to the rest of the structure in order to most effectively use the currently available knowledge. We show how this framework can be used also for detection rather than tracking which allows for automatic system initialization and for incorporating pose estimates obtained from independent object part detectors. Improved performance over alternative solutions is demonstrated on real-world sequences.

****

Occluding Contours for Multi-View Stereo

Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, Steven M. Seitz; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4002-4009

This paper leverages occluding contours (aka "internal silhouettes") to improve the performance of multi-view stereo methods. The contributions are 1) a new technique to identify free-space regions arising from occluding contours, and 2) a new approach for incorporating the resulting free-space constraints into Poisson surface reconstruction. The proposed approach outperforms state of the art MVS techniques for challenging Internet datasets, yielding dramatic quality improvements both around object contours and in surface detail.

****

Aerial Reconstructions via Probabilistic Data Fusion

Randi Cabezas, Oren Freifeld, Guy Rosman, John W. Fisher III; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4010-4017

We propose an integrated probabilistic model for multi-modal fusion of aerial imagery, LiDAR data, and (optional) GPS measurements. The model allows for analysis and dense reconstruction (in terms of both geometry and appearance) of large 3D scenes. An advantage of the approach is that it explicitly models uncertainty and allows for missing data. As compared with image-based methods, dense reconstructions of complex urban scenes are feasible with fewer observations. Moreover, the proposed model allows one to estimate absolute scale and orientation and reason about other aspects of the scene, e.g., detection of moving objects. As formulated, the model lends itself to massively-parallel computing. We exploit this in an efficient inference scheme that utilizes both general purpose and domain-specific hardware components. We demonstrate results on large-scale reconstruction of urban terrain from LiDAR and aerial photography data.

****

3D Modeling from Wide Baseline Range Scans using Contour Coherence

Ruizhe Wang, Jongmoo Choi, Gerard Medioni; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4018-4025

Registering 2 or more range scans is a fundamental problem, with application to 3D modeling. While this problem is well addressed by existing techniques such as ICP when the views overlap significantly at a good initialization, no satisfactory solution exists for wide baseline registration. We propose here a novel approach which leverages contour coherence and allows us to align two wide baseline range scans with limited overlap from a poor initialization. Inspired by ICP, we maximize the contour coherence by building robust corresponding pairs on apparent contours and minimizing their distances in an iterative fashion. We use the contour coherence under a multi-view rigid registration framework, and this enables the reconstruction of accurate and complete 3D models from as few as 4 frames. We further extend it to handle articulations, and this allows us to model articulated objects such as human body. Experimental results on both synthetic and r eal data demonstrate the effectiveness and robustness of our contour coherence based registration approach to wide baseline range scans, and to 3D modeling.

****

Ground Plane Estimation using a Hidden Markov Model

Ralf Dragon, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4026-4033

We focus on the problem of estimating the ground plane orientation and location in monocular video sequences from a moving observer. Our only assumptions are that the 3D ego motion t and the ground plane normal n are orthogonal, and that n and t are smooth over time. We formulate the problem as a state-continuous Hidden Markov Model (HMM) where the hidden state contains t and n and may be estimated by sampling and decomposing homographies. We show that using blocked Gibbs sampling, we can infer the hidden state with high robustness towards outliers, drifting trajectories, rolling shutter and an imprecise intrinsic calibration. Since our approach does not need any initial orientation prior, it works for arbitrary camera orientations in which the ground is visible.

****

Orientation Robust Text Line Detection in Natural Images

Le Kang, Yi Li, David Doermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4034-4041

In this paper, higher-order correlation clustering (HOCC) is used for text line detection in natural images. We treat text line detection as a graph partitioning problem, where each vertex is represented by a Maximally Stable Extremal Region (MSER). First, weak hypothesises are proposed by coarsely grouping MSERs based on their spatial alignment and appearance consistency. Then, higher-order correlation clustering (HOCC) is used to partition the MSERs into text line candidates, using the hypotheses as soft constraints to enforce long range interactions. We further propose a regularization method to solve the Semidefinite Programming problem in the inference. Finally we use a simple texton-based texture classifier to filter out the non-text areas. This framework allows us to naturally handle multiple orientations, languages and fonts. Experiments show that our approach achieves competitive performance compared to the state of the art.

****

Strokelets: A Learned Multi-Scale Representation for Scene Text Recognition

Cong Yao, Xiang Bai, Baoguang Shi, Wenyu Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4042-4049

Driven by the wide range of applications, scene text detection and recognition have become active research topics in computer vision. Though extensively studied, localizing and reading text in uncontrolled environments remain extremely challenging, due to various interference factors. In this paper, we propose a novel multi-scale representation for scene text recognition. This representation consists of a set of detectable primitives, termed as strokelets, which capture the essential substructures of characters at different granularities. Strokelets possess four distinctive advantages: (1) Usability: automatically learned from bounding box labels; (2) Robustness: insensitive to interference factors; (3) Generality: applicable to variant languages; and (4) Expressivity: effective at describing characters. Extensive experiments on standard benchmarks verify the advantages of strokelets and demonstrate the effectiveness of the proposed algorithm for text recognition.

****

Region-based Discriminative Feature Pooling for Scene Text Recognition

Chen-Yu Lee, Anurag Bhardwaj, Wei Di, Vignesh Jagadeesh, Robinson Piramuthu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4050-4057

We present a new feature representation method for scene text recognition problem, particularly focusing on improving scene character recognition. Many existing methods rely on Histogram of Oriented Gradient (HOG) or part-based models, which do not span the feature space well for characters in natural scene images, especially given large variation in fonts with cluttered backgrounds. In this work, we propose a discriminative feature pooling method that automatically learns th e most informative sub-regions of each scene character within a multi-class classification framework, whereas each sub-region seamlessly integrates a set of low-level image features through integral images. The proposed feature representation is compact, computationally efficient, and able to effectively model distinctive spatial structures of each individual character class. Extensive experiments conducted on challenging datasets (Chars74K, ICDAR'03, ICDAR'11, SVT) show that our method significantly outperforms existing methods on scene character classification and scene text recognition tasks.

****

Fast and Exact: ADMM-Based Discriminative Shape Segmentation with Loopy Part Models

Haithem Boussaid, Iasonas Kokkinos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4058-4065

In this work we use loopy part models to segment ensembles of organs in medical images. Each organ's shape is represented as a cyclic graph, while shape consistency is enforced through inter-shape connections. Our contributions are two-fold: firstly, we use an efficient decomposition-coordination algorithm to solve the resulting optimization problems: we decompose the model's graph into a set of open, chain-structured, graphs each of which is efficiently optimized using Dynamic Programming with Generalized Distance Transforms. We use the Alternating Direction Method of Multipliers (ADMM) to fix the potential inconsistencies of the individual solutions and show that ADMM yields substantially faster convergence than plain Dual Decomposition-based methods. Secondly, we employ structured prediction to encompass loss functions that better reflect the performance criteria used in medical image segmentation. By using the mean contour distance (MCD) as a structured loss during training, we obtain clear test-time performance gains. We demonstrate the merits of exact and efficient inference with rich, structured models in a large X-Ray image segmentation benchmark, where we obtain systematic improvements over the current state-of-the-art.

****

Pseudoconvex Proximal Splitting for L-infty Problems in Multiview Geometry

Anders Eriksson, Mats Isaksson; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4066-4073

In this paper we study optimization methods for minimizing large-scale pseudoconvex L_infinity problems in multiview geometry. We present a novel algorithm for solving this class of problem based on proximal splitting methods. We provide a brief derivation of the proposed method along with a general convergence analysis. The resulting meta-algorithm requires very little effort in terms of implementation and instead makes use of existing advanced solvers for non-linear optimization. Preliminary experiments on a number of real image datasets indicate that the proposed method experimentally matches or outperforms current state-of-the-art solvers for this class of problems.

****

A Convex Relaxation of the Ambrosio--Tortorelli Elliptic Functionals for the Mumford-Shah Functional

Youngwook Kee, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4074-4081

In this paper, we revisit the phase-field approximation of Ambrosio and Tortorelli for the Mumford--Shah functional. We then propose a convex relaxation for it to attempt to compute globally optimal solutions rather than solving the nonconvex functional directly, which is the main contribution of this paper. Inspired by McCormick's seminal work on factorable nonconvex problems, we split a nonconvex product term that appears in the Ambrosio--Tortorelli elliptic functionals in a way that a typical alternating gradient method guarantees a globally optimal solution without completely removing coupling effects. Furthermore, not only do we provide a fruitful analysis of the proposed relaxation but also demonstrate the capacity of our relaxation in numerous experiments that show convincing results compared to a naive extension of the McCormick relaxation and its quadratic variant. Indeed, we believe the proposed relaxation and the idea behind would open up a possibility for convexifying a new class of functions in the context of en

ergy minimization for computer vision.
****

Sequential Convex Relaxation for Mutual Information-Based Unsupervised Figure-Ground Segmentation

Youngwook Kee, Mohamed Souiai, Daniel Cremers, Junmo Kim; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4082-4089

We propose an optimization algorithm for mutual-information-based unsupervised figure-ground separation. The algorithm jointly estimates the color distributions of the foreground and background, and separates them based on their mutual information with geometric regularity. To this end, we revisit the notion of mutual information and reformulate it in terms of the photometric variable and the indicator function; and propose a sequential convex optimization strategy for solving the nonconvex optimization problem that arises. By minimizing a sequence of convex sub-problems for the mutual-information-based nonconvex energy, we efficiently attain high quality solutions for challenging unsupervised figure-ground segmentation problems. We demonstrate the capacity of our approach in numerous experiments that show convincing fully unsupervised figure-ground separation, in terms of both segmentation quality and robustness to initialization.

****

Decorrelated Vectorial Total Variation

Shunsuke Ono, Isao Yamada; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4090-4097

This paper proposes a new vectorial total variation prior (VTV) for color images. Different from existing VTVs, our VTV, named the decorrelated vectorial total variation prior (D-VTV), measures the discrete gradients of the luminance component and that of the chrominance one in a separated manner, which significantly reduces undesirable uneven color effects. Moreover, a higher-order generalization of the D-VTV, which we call the decorrelated vectorial total generalized variation prior (D-VTGV), is also developed for avoiding the staircasing effect that accompanies the use of VTVs. A noteworthy property of the D-VT(G)V is that it enables us to efficiently minimize objective functions involving it by a primal-dual splitting method. Experimental results illustrate their utility.

****

Efficient Squared Curvature

Claudia Nieuwenhuis, Eno Toeppe, Lena Gorelick, Olga Veksler, Yuri Boykov; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4098-4105

Curvature has received increasing attention as an important alternative to length based regularization in computer vision. In contrast to length, it preserves elongated structures and fine details. Existing approaches are either inefficient, or have low angular resolution and yield results with strong block artifacts. We derive a new model for computing squared curvature based on integral geometry. The model counts responses of straight line triple cliques. The corresponding energy decomposes into submodular and supermodular pairwise potentials. We show that this energy can be efficiently minimized even for high angular resolutions using the trust region framework. Our results confirm that we obtain accurate and visually pleasing solutions without strong artifacts at reasonable runtimes.

****

Multi-feature Spectral Clustering with Minimax Optimization

Hongxing Wang, Chaoqun Weng, Junsong Yuan; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4106-4113

In this paper, we propose a novel formulation for multi-feature clustering using minimax optimization. To find a consensus clustering result that is agreeable to all feature modalities, our objective is to find a universal feature embedding, which not only fits each individual feature modality well, but also unifies different feature modalities by minimizing their pairwise disagreements. The loss function consists of both (1) unary embedding cost for each modality, and (2) pairwise disagreement cost for each pair of modalities, with weighting parameters automatically selected to maximize the loss. By performing minimax optimization,

we can minimize the loss for the worst case with maximum disagreements, thus can better reconcile different feature modalities. To solve the minimax optimization, an iterative solution is proposed to update the universal embedding, individual embedding, and fusion weights, separately. Our minimax optimization has only one global parameter. The superior results on various multi-feature clustering tasks validate the effectiveness of our approach when compared with the state-of-the-art methods.
****
Quality-based Multimodal Classification using Tree-Structured Sparsity
Soheil Bahrampour, Asok Ray, Nasser M. Nasrabadi, Kenneth W. Jenkins; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4114-4121
Recent studies have demonstrated advantages of information fusion based on sparsity models for multimodal classification. Among several sparsity models, tree-structured sparsity provides a flexible framework for extraction of cross-correlated information from different sources and for enforcing group sparsity at multiple granularities. However, the existing algorithm only solves an approximated version of the cost functional and the resulting solution is not necessarily sparse at group levels. This paper reformulates the tree-structured sparse model for multimodal classification task. An accelerated proximal algorithm is proposed to solve the optimization problem, which is an efficient tool for feature-level fusion among either homogeneous or heterogeneous sources of information. In addition, a (fuzzy-set-theoretic) possibilistic scheme is proposed to weight the available modalities, based on their respective reliability, in a joint optimization problem for finding the sparsity codes. This approach provides a general framework for quality-based fusion that offers added robustness to several sparsity-based multimodal classification algorithms. To demonstrate their efficacy, the proposed methods are evaluated on three different applications - multiview face recognition, multimodal face recognition, and target classification.
****
Newton Greedy Pursuit: A Quadratic Approximation Method for Sparsity-Constrained Optimization
Xiao-Tong Yuan, Qingshan Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4122-4129
First-order greedy selection algorithms have been widely applied to sparsity-constrained optimization. The main theme of this type of methods is to evaluate the function gradient in the previous iteration to update the non-zero entries and their values in the next iteration. In contrast, relatively less effort has been made to study the second-order greedy selection method additionally utilizing the Hessian information. Inspired by the classic constrained Newton method, we propose in this paper the NewTon Greedy Pursuit (NTGP) method to approximately minimizes a twice differentiable function over sparsity constraint. At each iteration, NTGP constructs a second-order Taylor expansion to approximate the cost function, and estimates the next iterate as the solution of the constructed quadratic model over sparsity constraint. Parameter estimation error and convergence property of NTGP are analyzed. The superiority of NTGP to several representative first-order greedy selection methods is demonstrated in synthetic and real sparse logistic regression tasks.
****
Generalized Nonconvex Nonsmooth Low-Rank Minimization
Canyi Lu, Jinhui Tang, Shuicheng Yan, Zhouchen Lin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4130-4137
As surrogate functions of L_0-norm, many nonconvex penalty functions have been proposed to enhance the sparse vector recovery. It is easy to extend these nonconvex penalty functions on singular values of a matrix to enhance low-rank matrix recovery. However, different from convex optimization, solving the nonconvex low-rank minimization problem is much more challenging than the nonconvex sparse minimization problem. We observe that all the existing nonconvex penalty functions are concave and monotonically increasing on [0,\infty). Thus their gradients are decreasing functions. Based on this property, we propose an Iteratively Reweig

hted Nuclear Norm (IRNN) algorithm to solve the nonconvex nonsmooth low-rank minimization problem. IRNN iteratively solves a Weighted Singular Value Thresholding (WSVT) problem. By setting the weight vector as the gradient of the concave penalty function, the WSVT problem has a closed form solution. In theory, we prove that IRNN decreases the objective function value monotonically, and any limit point is a stationary point. Extensive experiments on both synthetic data and real images demonstrate that IRNN enhances the low-rank matrix recovery compared with state-of-the-art convex algorithms.

****

Latent Dictionary Learning for Sparse Representation based Classification

Meng Yang, Dengxin Dai, Lilin Shen, Luc Van Gool; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4138-4145

Dictionary learning (DL) for sparse coding has shown promising results in classification tasks, while how to adaptively build the relationship between dictionary atoms and class labels is still an important open question. The existing dictionary learning approaches simply fix a dictionary atom to be either class-specific or shared by all classes beforehand, ignoring that the relationship needs to be updated during DL. To address this issue, in this paper we propose a novel latent dictionary learning (LDL) method to learn a discriminative dictionary and build its relationship to class labels adaptively. Each dictionary atom is jointly learned with a latent vector, which associates this atom to the representation of different classes. More specifically, we introduce a latent representation model, in which discrimination of the learned dictionary is exploited via minimizing the within-class scatter of coding coefficients and the latent-value weighted dictionary coherence. The optimal solution is efficiently obtained by the proposed solving algorithm. Correspondingly, a latent sparse representation based classifier is also presented. Experimental results demonstrate that our algorithm outperforms many recently proposed sparse representation and dictionary learning approaches for action, gender and face recognition.

****

Is Rotation a Nuisance in Shape Recognition?

Qiuhong Ke, Yi Li; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4146-4153

Rotation in closed contour recognition is a puzzling nuisance in most algorithms. In this paper we address three fundamental issues brought by rotation in shapes: 1) is alignment among shapes necessary? If the answer is "no", 2) how to exploit information in different rotations? and 3) how to use rotation unaware local features for rotation aware shape recognition? We argue that the origin of these issues is the use of hand crafted rotation-unfriendly features and measurements. Therefore our goal is to learn a set of hierarchical features that describe all rotated versions of a shape as a class, with the capability of distinguishing different such classes. We propose to rotate shapes as many times as possible as training samples, and learn the hierarchical feature representation by effectively adopting a convolutional neural network. We further show that our method is very efficient because the network responses of all possible shifted versions of the same shape can be computed effectively by re-using information in the overlapping areas. We tested the algorithm on three real datasets: Swedish Leaves dataset, ETH-80 Shape, and a subset of the recently collected Leafsnap dataset. Our approach used the curvature scale space and outperformed the state of the art.

****

Dual-Space Decomposition of 2D Complex Shapes

Guilin Liu, Zhonghua Xi, Jyh-Ming Lien; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4154-4161

While techniques that segment shapes into visually meaningful parts have generated impressive results, these techniques also have only focused on relatively simple shapes, such as those composed of a single object either without holes or with few simple holes. In many applications, shapes created from images can contain many overlapping objects and holes. These holes may come from sensor noise, ma y have important parts of the shape or may be arbitrarily complex. These complexities that appear in real-world 2D shapes can pose grand challenges to the existing part segmentation methods. In this paper, we propose a new decomposition method, called Dual-space Decomposition that handles complex 2D shapes by recognizing the importance of holes and classifying holes as either topological noise or structurally important features. Our method creates a nearly convex decomposition of a given shape by segmenting both the polygon itself and its complementary. We compare our results to segmentation produced by nonexpert human subjects. Based on two evaluation methods, we show that this new decomposition method creates statistically similar to those produced by human subjects.

****

Noising versus Smoothing for Vertex Identification in Unknown Shapes

Konstantinos A. Raftopoulos, Marin Ferecatu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4162-4168

A method for identifying shape features of local nature on the shape's boundary, in a way that is facilitated by the presence of noise is presented. The boundary is seen as a real function. A study of a certain distance function reveals, almost counter-intuitively, that vertices can be defined and localized better in the presence of noise, thus the concept of noising, as opposed to smoothing, is conceived and presented. The method works on both smooth and noisy shapes, the presence of noise having an effect of improving on the results of the smoothed version. Experiments with noise and a comparison to state of the art validate the method.

****

Surface Registration by Optimization in Constrained Diffeomorphism Space

Wei Zeng, Lok Ming Lui, Xianfeng Gu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4169-4176

This work proposes a novel framework for optimization in the constrained diffeomorphism space for deformable surface registration. First the diffeomorphism space is modeled as a special complex functional space on the source surface, the Beltrami coefficient space. The physically plausible constraints, in terms of feature landmarks and deformation types, define subspaces in the Beltrami coefficient space. Then the harmonic energy of the registration is minimized in the constrained subspaces. The minimization is achieved by alternating two steps: 1) optimization - diffuse the Beltrami coefficient, and 2) projection - first deform the conformal structure by the current Beltrami coefficient and then compose with a harmonic map from the deformed conformal structure to the target. The registration result is diffeomorphic, satisfies the physical landmark and deformation constraints, and minimizes the conformality distortion. Experiments on human facial surfaces demonstrate the efficiency and efficacy of the proposed registration framework.

****

Dense Non-Rigid Shape Correspondence using Random Forests

Emanuele Rodola, Samuel Rota Bulo, Thomas Windheuser, Matthias Vestner, Daniel Cremers; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4177-4184

We propose a shape matching method that produces dense correspondences tuned to a specific class of shapes and deformations. In a scenario where this class is represented by a small set of example shapes, the proposed method learns a shape descriptor capturing the variability of the deformations in the given class. The approach enables the wave kernel signature to extend the class of recognized deformations from near isometries to the deformations appearing in the example set by means of a random forest classifier. With the help of the introduced spatial regularization, the proposed method achieves significant improvements over the baseline approach and obtains state-of-the-art results while keeping short computation times.

****

Covariance Descriptors for 3D Shape Matching and Retrieval

Hedi Tabia, Hamid Laga, David Picard, Philippe-Henri Gosselin; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4

185-4192

Several descriptors have been proposed in the past for 3D shape analysis, yet none of them achieves best performance on all shape classes. In this paper we propose a novel method for 3D shape analysis using the covariance matrices of the descriptors rather than the descriptors themselves. Covariance matrices enable efficient fusion of different types of features and modalities. They capture, using the same representation, not only the geometric and the spatial properties of a shape region but also the correlation of these properties within the region. Covariance matrices, however, lie on the manifold of Symmetric Positive Definite (SPD) tensors, a special type of Riemannian manifolds, which makes comparison and clustering of such matrices challenging. In this paper we study covariance matrices in their native space and make use of geodesic distances on the manifold as a dissimilarity measure. We demonstrate the performance of this metric on 3D face matching and recognition tasks. We then generalize the Bag of Features paradigm, originally designed in Euclidean spaces, to the Riemannian manifold of SPD matrices. We propose a new clustering procedure that takes into account the geometry of the Riemannian manifold. We evaluate the performance of the proposed Bag of Covariance Matrices framework on 3D shape matching and retrieval applications and demonstrate its superiority compared to descriptor-based techniques.

****

Symmetry-Aware Nonrigid Matching of Incomplete 3D Surfaces

Yusuke Yoshiyasu, Eiichi Yoshida, Kazuhito Yokoi, Ryusuke Sagawa; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4193-4200

We present a nonrigid shape matching technique for establishing correspondences of incomplete 3D surfaces that exhibit intrinsic reflectional symmetry. The key for solving the symmetry ambiguity problem is to use a point-wise local mesh descriptor that has orientation and is thus sensitive to local reflectional symmetry, e.g. discriminating the left hand and the right hand. We devise a way to compute the descriptor orientation by taking the gradients of a scalar field called the average diffusion distance (ADD). Because ADD is smoothly defined on a surface, invariant under isometry/scale and robust to topological errors, the robustness of the descriptor to non-rigid deformations is improved. In addition, we propose a graph matching algorithm called iterative spectral relaxation which combines spectral embedding and spectral graph matching. This formulation allows us to define pairwise constraints in a scale-invariant manner from k-nearest neighbor local pairs such that non-isometric deformations can be robustly handled. Experimental results show that our method can match challenging surfaces with global intrinsic symmetry, data incompleteness and non-isometric deformations.

****

An Automated Estimator of Image Visual Realism Based on Human Cognition

Shaojing Fan, Tian-Tsong Ng, Jonathan S. Herberg, Bryan L. Koenig, Cheston Y.-C. Tan, Rangding Wang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4201-4208

Assessing the visual realism of images is increasingly becoming an essential aspect of fields ranging from computer graphics (CG) rendering to photo manipulation. In this paper we systematically evaluate factors underlying human perception of visual realism and use that information to create an automated assessment of visual realism. We make the following unique contributions. First, we established a benchmark dataset of images with empirically determined visual realism scores. Second, we identified attributes potentially related to image realism, and used correlational techniques to determine that realism was most related to image naturalness, familiarity, aesthetics, and semantics. Third, we created an attributes-motivated, automated computational model that estimated image visual realism quantitatively. Using human assessment as a benchmark, the model was below human performance, but outperformed other state-of-the-art algorithms.

****

SteadyFlow: Spatially Smooth Optical Flow for Video Stabilization

Shuaicheng Liu, Lu Yuan, Ping Tan, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4209-4216

We propose a novel motion model, SteadyFlow, to represent the motion between neighboring video frames for stabilization. A SteadyFlow is a specific optical flow by enforcing strong spatial coherence, such that smoothing feature trajectories can be replaced by smoothing pixel profiles, which are motion vectors collected at the same pixel location in the SteadyFlow over time. In this way, we can avoid brittle feature tracking in a video stabilization system. Besides, SteadyFlow is a more general 2D motion model which can deal with spatially-variant motion. We initialize the SteadyFlow by optical flow and then discard discontinuous motions by a spatial-temporal analysis and fill in missing regions by motion completion. Our experiments demonstrate the effectiveness of our stabilization on real-world challenging videos.

****

Automatic Face Reenactment

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, Christian Theobalt; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4217-4224

We propose an image-based, facial reenactment system that replaces the face of an actor in an existing target video with the face of a user from a source video, while preserving the original target performance. Our system is fully automatic and does not require a database of source expressions. Instead, it is able to produce convincing reenactment results from a short source video captured with an off-the-shelf camera, such as a webcam, where the user performs arbitrary facial gestures. Our reenactment pipeline is conceived as part image retrieval and part face transfer: The image retrieval is based on temporal clustering of target frames and a novel image matching metric that combines appearance and motion to select candidate frames from the source video, while the face transfer uses a 2D warping strategy that preserves the user's identity. Our system excels in simplicity as it does not rely on a 3D face model, it is robust under head motion and does not require the source and target performance to be similar. We show convincing reenactment results for videos that we recorded ourselves and for low-quality footage taken from the Internet.

****

Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction

Gunhee Kim, Leonid Sigal, Eric P. Xing; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4225-4232

In this paper, we address the problem of jointly summarizing large sets of Flickr images and YouTube videos. Starting from the intuition that the characteristics of the two media types are different yet complementary, we develop a fast and easily-parallelizable approach for creating not only high-quality video summaries but also novel structural summaries of online images as storyline graphs. The storyline graphs can illustrate various events or activities associated with the topic in a form of a branching network. The video summarization is achieved by diversity ranking on the similarity graphs between images and video frames. The reconstruction of storyline graphs is formulated as the inference of sparse time-varying directed graphs from a set of photo streams with assistance of videos. For evaluation, we collect the datasets of 20 outdoor activities, consisting of 2.7M Flickr images and 16K YouTube videos. Due to the large-scale nature of our problem, we evaluate our algorithm via crowdsourcing using Amazon Mechanical Turk. In our experiments, we demonstrate that the proposed joint summarization approach outperforms other baselines and our own methods using videos or images only.

****

Semi-supervised Relational Topic Model for Weakly Annotated Image Recognition in Social Media

Zhenxing Niu, Gang Hua, Xinbo Gao, Qi Tian; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4233-4240

In this paper, we address the problem of recognizing images with weakly annotated text tags. Most previous work either cannot be applied to the scenarios where the tags are loosely related to the images; or simply take a pre-fusion at the f eature level or a post-fusion at the decision level to combine the visual and textual content. Instead, we first encode the text tags as the relations among the images, and then propose a semi-supervised relational topic model (ss-RTM) to explicitly model the image content and their relations. In such way, we can efficiently leverage the loosely related tags, and build an intermediate level representation for a collection of weakly annotated images. The intermediate level representation can be regarded as a mid-level fusion of the visual and textual content, which is able to explicitly model their intrinsic relationships. Moreover, image category labels are also modeled in the ss-RTM, and recognition can be conducted without training an additional discriminative classifier. Our extensive experiments on social multimedia datasets (images+tags) demonstrated the advantages of the proposed model.

****

Beyond Human Opinion Scores: Blind Image Quality Assessment based on Synthetic Scores

Peng Ye, Jayant Kumar, David Doermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4241-4248

State-of-the-art general purpose Blind Image Quality Assessment (BIQA) models rely on examples of distorted images and corresponding human opinion scores to learn a regression function that maps image features to a quality score. These types of models are considered "opinion-aware" (OA) BIQA models. A large set of human scored training examples is usually required to train a reliable OA-BIQA model. However, obtaining human opinion scores through subjective testing is often expensive and time-consuming. It is therefore desirable to develop "opinion-free" (OF) BIQA models that do not require human opinion scores for training. This paper proposes BLISS (Blind Learning of Image Quality using Synthetic Scores). BLISS is a simple, yet effective method for extending OA-BIQA models to OF-BIQA models. Instead of training on human opinion scores, we propose to train BIQA models on synthetic scores derived from Full-Reference (FR) IQA measures. State-of-the-art FR measures yield high correlation with human opinion scores and can serve as approximations to human opinion scores. Unsupervised rank aggregation is applied to combine different FR measures to generate a synthetic score, which serves as a better "gold standard". Extensive experiments on standard IQA datasets show that BLISS significantly outperforms previous OF-BIQA methods and is comparable to state-of-the-art OA-BIQA methods.

****

Active Sampling for Subjective Image Quality Assessment

Peng Ye, David Doermann; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4249-4256

Subjective Image Quality Assessment (IQA) is the most reliable way to evaluate the visual quality of digital images perceived by the end user. It is often used to construct image quality datasets and provide the groundtruth for building and evaluating objective quality measures. Subjective tests based on the Mean Opinion Score (MOS) have been widely used in previous studies, but have many known problems such as an ambiguous scale definition and dissimilar interpretations of the scale among subjects. To overcome these limitations, Paired Comparison (PC) tests have been proposed as an alternative and are expected to yield more reliable results. However, PC tests can be expensive and time consuming, since for n images they require n choose 2 comparisons. We present a hybrid subjective test which combines MOS and PC tests via a unified probabilistic model and an active sampling method. The proposed method actively constructs a set of queries consisting of MOS and PC tests based on the expected information gain provided by each test and can effectively reduce the number of tests required for achieving a target accuracy. Our method can be used in conventional laboratory studies as well as crowdsourcing experiments. Experimental results show the proposed method outperforms state-of-the-art subjective IQA tests in a crowdsourced setting.

****

A Study on Cross-Population Age Estimation

Guodong Guo, Chao Zhang; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4257-4263

We study the problem of cross-population age estimation. Human aging is determined by the genes and influenced by many factors. Different populations, e.g., males and females, Caucasian and Asian, may age differently. Previous research has discovered the aging difference among different populations, and reported large errors in age estimation when crossing gender and/or ethnicity. In this paper we propose novel methods for cross-population age estimation with a good performance. The proposed methods are based on projecting the different aging patterns into a common space where the aging patterns can be correlated even though they come from different populations. The projections are also discriminative between age classes due to the integration of the classical discriminant analysis technique. Further, we study the amount of data needed in the target population to learn a cross-population age estimator. Finally, we study the feasibility of multi-source cross-population age estimation. Experiments are conducted on a large database of more than 21,000 face images selected from the MORPH. Our studies are valuable to significantly reduce the burden of training data collection for age estimation on a new population, utilizing existing aging patterns even from different populations.

****

Remote Heart Rate Measurement From Face Videos Under Realistic Situations

Xiaobai Li, Jie Chen, Guoying Zhao, Matti Pietikainen; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4264-4271

Heart rate is an important indicator of people's physiological state. Recently, several papers reported methods to measure heart rate remotely from face videos. Those methods work well on stationary subjects under well controlled conditions, but their performance significantly degrades if the videos are recorded under more challenging conditions, specifically when subjects' motions and illumination variations are involved. We propose a framework which utilizes face tracking and Normalized Least Mean Square adaptive filtering methods to counter their influences. We test our framework on a large difficult and public database MAHNOB-HCI and demonstrate that our method substantially outperforms all previous methods. We also use our method for long term heart rate monitoring in a game evaluation scenario and achieve promising results.

****

6 Seconds of Sound and Vision: Creativity in Micro-Videos

Miriam Redi, Neil O'Hare, Rossano Schifanella, Michele Trevisiol, Alejandro Jaimes; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4272-4279

The notion of creativity, as opposed to related concepts such as beauty or interestingness, has not been studied from the perspective of automatic analysis of multimedia content. Meanwhile, short online videos shared on social media platforms, or micro-videos, have arisen as a new medium for creative expression. In this paper we study creative micro-videos in an effort to understand the features that make a video creative, and to address the problem of automatic detection of creative content. Defining creative videos as those that are novel and have aesthetic value, we conduct a crowdsourcing experiment to create a dataset of over 3,800 micro-videos labelled as creative and non-creative. We propose a set of computational features that we map to the components of our definition of creativity, and conduct an analysis to determine which of these features correlate most with creative video. Finally, we evaluate a supervised approach to automatically detect creative video, with promising results, showing that it is necessary to model both aesthetic value and novelty to achieve optimal classification accuracy.

****

GPS-Tag Refinement using Random Walks with an Adaptive Damping Factor

Amir Roshan Zamir, Shervin Ardeshir, Mubarak Shah; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4280-4287

The number of GPS-tagged images available on the web is increasing at a rapid rate. The majority of such location tags are specified by the users, either through manual tagging or localization-chips embedded in the cameras. However, a known issue with user shared images is the unreliability of such GPS-tags. In this pa per, we propose a method for addressing this problem. We assume a large dataset of GPS-tagged images which includes an unknown subset with contaminated tags is available. We develop a robust method for identification and refinement of this subset using the rest of the images in the dataset. In the proposed method, we form a large number of triplets of matching images and use them for estimating the location of the query image utilizing structure from motion. Some of the generated estimations may be inaccurate due to the noisy GPS-tags in the dataset. Therefore, we perform Random Walks on the estimations in order to identify the subset with the maximal agreement. Finally, we estimate the GPS-tag of the query utilizing the identified consistent subset using a weighted mean. We propose a new damping factor for Random Walks which conforms to the level of noise in the input, and consequently, robustifies Random Walks. We evaluated the proposed framework on a dataset of over 18k user-shared images; the experiments show our method robustly improves the accuracy of GPS-tags under diverse scenarios.

****